



新世纪高等学校教材

环境科学与工程系列教材

北京师范大学环境学院 组编

环境统计 分析

杨晓华 刘瑞民 曾 勇 编著

HUANJI
TONGJI
FENXI



北京师范大学出版集团
BEIJING NORMAL UNIVERSITY PUBLISHING GROUP
北京师范大学出版社



新世纪高等学校教材

环境科学与工程系列教材

湿地学

环境水力学原理

环境影响评价实用教程

环境科学案例研究

环境科学案例研究教师手册

城市生态规划学

城市空气质量规划与管理

■ 环境统计分析

环境与健康

ISBN 978-7-303-09502-5



9 787303 095025 >

定价: 35.00 元

新世纪高等学校教材

环境科学与工程系列教材

北京师范大学环境学院 组编

环境统计分析

HUANJIANG TONGJI FENXI



杨晓华 刘瑞民 曾勇 编著



北京师范大学出版集团

北京师范大学出版社

图书在版编目(CIP)数据

环境统计分析/杨晓华,刘瑞民,曾勇编著. —北京:北京师范大学出版社, 2008.8

(环境科学与工程系列教材)

ISBN 978-7-303-09502-5

I. 环… II. ①杨… ②刘… ③曾… III. 环境统计-统计分析-高等学校-教材 IV. X11

中国版本图书馆 CIP 数据核字(2008)第 113056 号

出版发行: 北京师范大学出版社 www.bnup.com.cn

北京新街口外大街 19 号

邮政编码: 100875

印 刷: 北京新丰印刷厂

经 销: 全国新华书店

开 本: 170mm×230 mm

印 张: 20.25

字 数: 337 千字

印 数: 1~3 000 册

版 次: 2008 年 9 月第 1 版

印 次: 2008 年 9 月第 1 次印刷

定 价: 35.00 元

责任编辑: 毛 佳

装帧设计: 高 霞

责任校对: 李 蕊

责任印制: 马鸿麟

版权所有 侵权必究

反盗版、侵权举报电话: 010-58800697

北京读者服务部电话: 010-58808104

外埠邮购电话: 010-58808083

本书如有印装质量问题, 请与印制管理部联系调换。

印制管理部电话: 010-58800825

前

言

环境统计分析是环境科学与环境工程的基础学科之一,是一门对环境系统不确定性问题进行数据处理、模型构建和分析的学科。环境系统,系指地球表面包括非生物、生物的各种环境因素及其相互关系的总和,是一个具有时、空、量、序变化的复杂巨系统。受人类活动、天文、气候和气象等众多因素的影响,环境系统中存在许多不确定性现象,并且有大量的数据需要进行统计分析和处理。环境的理论和实践对统计信息的需求急剧增加,对统计分析的理论和方法提出了更高的要求。在自然、社会与环境关系的基础上,用统计方法对环境问题予以量化描述和分析已成为环境研究的迫切需要。环境统计学的产生与发展使人们能够利用数理统计方法处理或解决环境中的不确定性问题,使其定量化,其中包括寻找变量之间的定量关系、从数据中发现环境趋势、探索环境系统变化规律。现代环境统计学一个很重要的标志就是模型技术的运用及量化分析。

全书分三大部分,共10章。其中,第1章属于基础篇,简要地介绍了环境统计分析的概率统计基础知识;第2~9章属于模型篇,阐述了环境一元线性回归分析、环境多元线性回归分析、环境系统聚类分析、环境模糊聚类分析、环境判别分析、环境主成分分析、环境因子分析、人工神经网络等方法、模型的原理,并给出了分析案例;第10章属于空间分析篇,介绍了环境空间统计分析的基本原理,

并给出了应用实例。全书的大多数例子都是用目前常用的统计分析语言 Matlab 编写实现的,是理论联系实际的经验总结,具有可操作性。本书适于做高等院校环境科学与环境工程专业的高年级本科生和研究生教材,对环境科学与环境工程、生态学、资源与管理、应用数学、地理科学等相关领域的学者和科研人员也有重要的参考价值。

本书第1章由杨晓华、曾勇执笔,第2~9章由杨晓华执笔,第10章由刘瑞民执笔,全书由杨晓华统稿。另外,尹心安参加了第1章的编写工作;王伟参加了第3章、第4章、第10章的编写工作;陈强、胡晓雪参加了第5章、第6章的编写工作;余敦先参加了第1章、第3章、第6章、第8章的编写工作。2004级、2005级的博士研究生、2005级的硕士研究生也提供了部分例题和习题。另外,习题答案均是用 Matlab 语言计算完成。

在本书的编写和出版过程中,北京师范大学环境学院院长杨志峰教授,副院长沈珍瑶、刘静玲教授,还有牛军峰、孙涛副教授以及北京师范大学出版社的胡廷兰、毛佳等同志对本书提出了许多宝贵意见。书中若干例题选自所列参考文献,在此一并表示感谢。由于我们的水平有限,书中错误在所难免,欢迎读者批评指正。

衷心感谢北京师范大学出版社给予的大力支持!

本书的完成得到国家重点基础研究发展规划项目(G2003CB415204)的资助,在此表示衷心的感谢!

编著者

2007年7月

内 容 提 要

本书阐述了常用的环境统计分析方法，并给出了分析案例。首先简明扼要地介绍了环境统计分析的概率统计基础知识，又重点阐述了环境一元线性回归分析、环境多元线性回归分析、环境系统聚类分析、环境模糊聚类分析、环境判别分析、环境主成分分析和环境因子分析这些常用的环境统计分析模型；另外还给出了现代环境数据处理常用的人工神经网络方法和空间统计分析方法。对每一种方法，本书除了讲明基本原理外，还给出了大量的计算分析例题和案例。本书的部分例子是用目前实用的统计分析语言 Matlab 编写实现的，是理论联系实际的经验总结，具有实用性。本书适于做高等院校环境科学与环境工程专业的高年级本科生和研究生教材，对环境科学与环境工程、生态学、资源与管理、应用数学、地理科学等相关领域的学者和科研人员也有重要的参考价值。

目 录

第 1 章 概率统计基础 (1)

1.1 四种重要的概率分布 (1)

1.1.1 正态分布 (1)

1.1.2 χ^2 分布 (4)

1.1.3 t 分布 (5)

1.1.4 F 分布 (6)

1.2 随机向量的数字特征 (7)

1.2.1 数学期望 (7)

1.2.2 方差和均方差 (10)

1.2.3 原点矩和中心矩 (11)

1.2.4 变异系数 (12)

1.2.5 协方差阵和自协方差阵 (12)

1.2.6 随机变量的相关系数 (13)

1.2.7 总体与样本 (15)

1.2.8 样本子样的一些数字特征 (16)

1.2.9 大数定律 (16)

1.2.10 中心极限定理 (18)

1.3 参数估计 (20)

1.3.1 点估计 (21)

1.3.2 区间估计 (21)

1.4 参数假设检验 (24)

1.4.1 假设检验的原理 (25)

1.4.2	假设检验的步骤	(26)
1.4.3	参数检验	(27)
1.5	方差分析与试验设计初步	(34)
1.5.1	方差分析概述	(34)
1.5.2	单因素方差分析	(35)
1.5.3	双因素方差分析	(39)
1.5.4	试验设计初步	(45)
思考题1		(48)
参考文献		(49)

第2章 环境一元线性回归分析 (50)

2.1	一元线性回归模型	(50)
2.1.1	变量间的统计关系	(50)
2.1.2	一元线性回归模型	(52)
2.1.3	最小二乘法估计	(54)
2.2	线性回归方程的显著性检验	(55)
2.2.1	F 检验法	(56)
2.2.2	相关系数检验法	(58)
2.2.3	样本决定系数 r^2	(59)
2.3	线性回归式的误差估计	(60)
2.3.1	线性回归式的误差估计	(60)
2.3.2	线性回归的步骤	(61)
2.4	可化为一元线性回归的曲线回归	(62)
2.4.1	倒数变换	(62)
2.4.2	对数变换	(63)
2.4.3	混合变换	(64)
2.5	环境应用	(65)
思考题2		(69)
参考文献		(70)

第3章 环境多元线性回归分析 (71)

3.1	多元线性回归模型	(71)
3.2	参数的最小二乘估计	(72)

3.3 回归方程的显著性检验	(74)
3.3.1 拟合优度检验	(75)
3.3.2 F 检验	(76)
3.4 回归系数的显著性检验	(77)
3.5 Matlab 语言在多元回归中的应用	(79)
3.6 环境应用	(81)
思考题 3	(84)
参考文献	(86)

第 4 章 环境系统聚类分析 (87)

4.1 聚类分析概述	(87)
4.2 聚类要素的数据处理	(88)
4.3 距离和相似系数的计算	(93)
4.3.1 距离的计算	(93)
4.3.2 相似系数的计算	(97)
4.3.3 距离和相似系数选择原则	(99)
4.4 系统聚类分析常用方法	(100)
4.4.1 最短距离系统聚类法原理	(102)
4.4.2 最远距离聚类法原理	(103)
4.4.3 系统聚类法公式的统一	(105)
4.5 环境应用	(107)
思考题 4	(112)
参考文献	(115)

第 5 章 环境模糊聚类分析 (116)

5.1 模糊集理论	(116)
5.1.1 模糊集的基本概念	(117)
5.1.2 模糊集表示方法	(117)
5.1.3 模糊集的运算	(119)
5.1.4 模糊映射	(120)
5.2 模糊关系	(120)
5.3 模糊等价关系	(121)
5.4 模糊聚类分析步骤	(123)

5.4.1 数据标准化	110
5.4.2 模糊相似矩阵的建立	111
聚类分析	112
5.4.4 分类的 F 检验	113
5.5 环境应用	114
思考题 5	115
参考文献	116

第 6 章 环境判别分析 (119)

6.1 距离判别分析	119
6.1.1 两总体情况	119
6.1.2 多总体情况	120
6.2 Fisher 判别	121
6.3 Bayes 判别	122
6.4 环境应用	123
思考题 6	124
参考文献	125

第 7 章 环境主成分分析 (161)

7.1 主成分分析概述	161
7.2 主成分分析计算原理	165
7.3 主成分分析的性质	169
7.4 环境应用	170
思考题 7	178
参考文献	180

第 8 章 环境因子分析 (181)

因子分析概述	181
正交因子模型	182
正交因子模型的统计意义	184
正交因子模型的求解	185
因子旋转	188

8.6 因子得分	(191)
8.7 环境应用	(193)
思考题 8 ..	1
参考文献 ..	

第 9 章 人工神经网络 (206)

9.1 人工神经网络概述	
9.2 人工神经元模型	
9.3 BP 神经网络	
9.3.1 BP 神经网络原理	
9.3.2 BP 算法	
9.3.3 环境应用	
9.4 RBF 神经网络	
9.4.1 RBF 神经网络原理	
9.4.2 RBF 神经网络模型	
9.4.3 环境应用	
思考题	
参考文献	

第 10 章 环境空间统计分析 (232)

10.1 环境空间信息概述	
10.1.1 环境空间信息特征	
10.1.2 环境空间信息种类	
10.1.3 环境空间信息来源	
10.2 环境空间统计分析	
10.2.1 区域化变量	
10.2.2 协方差函数	(238)
10.2.3 变差函数	(239)
10.2.4 普通克立格插值	(248)
10.2.5 环境应用	(252)
10.3 环境空间主成分分析	
10.3.1 空间主成分分析步骤	
10.3.2 环境应用	

参考文献	11
参考文献	11

部分思考题答案	25
---------	----

附录	25
----	----

附表1 标准正态分布表	25
附表2 相关系数检验表	25
附表3 χ^2 分布临界值表	25
附表4 t 分布临界值表	25
附表5 F 分布临界值表	25

第1章 概率统计基础

统计与统计模型是环境科学中应用最为广泛的知识,对统计模型的理论和方法提出更高的要求。在自然科学和社会科学、工程的基础理论、应用方法与环境科学中,统计与统计模型是环境科学中应用最为广泛的知识,对统计模型的理论和方法提出更高的要求。在自然科学和社会科学、工程的基础理论、应用方法与环境科学中,统计与统计模型是环境科学中应用最为广泛的知识,对统计模型的理论和方法提出更高的要求。

本章的主要内容是:

- 四种重要的概率分布;
- 随机向量的数字特征;
- 参数估计;
- 参数假设检验;
- 方差分析与试验设计初步

1.1 四种重要的概率分布

在环境科学中,常用的四种重要的概率分布是:正态分布、泊松分布、二项分布、指数分布。正态分布、泊松分布、二项分布、指数分布是环境科学中应用最为广泛的知识,对统计模型的理论和方法提出更高的要求。

1.1.1 正态分布

在环境科学中,常用的四种重要的概率分布是:正态分布、泊松分布、二项分布、指数分布。正态分布、泊松分布、二项分布、指数分布是环境科学中应用最为广泛的知识,对统计模型的理论和方法提出更高的要求。

因此, 当人数足够多时, 总体中符合一定条件者, 许多不是正态分布的样本均值在样本量很大时, 也可用正态分布来近似。

若随机变量 X 的分布密度为:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (-\infty < x < +\infty, \sigma > 0) \quad (1.1)$$

则称 X 服从正态分布 $N(\mu, \sigma^2)$, 记为 $X \sim N(\mu, \sigma^2)$ 。其中, μ 为均值, σ 为标准差, σ^2 为方差 (标准差的平方)。

正态分布与 χ^2 分布、 F 分布、 t 分布、泊松分布一起是统计推断中最重要的分布, 也是分布, 各统计量分布都服从正态分布, 所以统计推断中不可少了正态分布, 且正态分布是其他分布的极限分布, 如 χ^2 分布、 F 分布、 t 分布、泊松分布 (Poisson distribution)。标准正态分布的密度函数与分布函数记为:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad (-\infty < x < +\infty) \quad (1.2)$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{y^2}{2}} dy \quad (-\infty < x < +\infty) \quad (1.3)$$

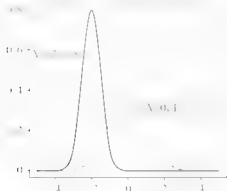


图 1-1 两条正态分布的密度曲线图
(左边是 $N(-2, 0.5)$ 分布, 右边是 $N(0, 1)$ 分布)

在统计推断中, 我们常会遇到总体服从正态分布的情况, 现将一般正态分布标准化, 使之成为标准正态分布, 即服从正态分布的标准化过程。

设 $X \sim N(\mu, \sigma^2)$, 作线性变换, 减去其均值, 再除以标准差 σ , 则很容易得到随机变量 $Y = \frac{X - \mu}{\sigma} \sim N(0, 1)$ 。

因为:

$$E(Y) = E\left(\frac{X - \mu}{\sigma}\right) = \frac{1}{\sigma} E(X - \mu)$$

$$D(Y) = D\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma^2} D(X) = 1$$

这样就将一般正态分布变换成了标准正态分布

由于 $\frac{X-\mu}{\sigma}$ 在正态分布中服从标准正态分布, 故是 $F(x)$ 。为简便, 今后记 $F(x)$ 为标准正态分布函数, 本书凡提到分布函数时又都写 $F(x)$ 。

设 $X \sim N(0, 1)$, 若 z_α 满足条件

$$P(X > z_\alpha) = \alpha \quad (0 < \alpha < 1)$$

则称 z_α 为标准正态分布的上侧 α 分位点, 记为 z_α 。如图 1-2 所示。



图 1-2 标准正态分布的上侧 α 分位点 z_α

例如, 查附表 1, 知 $z_{0.154} = 1.019\ 428$, 即 $P(X > 1.019\ 428) = 0.154$ 。

例 1.1 已知 $X \sim N(\mu, \sigma^2)$, X 在区间 $(\mu - k\sigma, \mu + k\sigma)$ 内取值, 求 $k = 1, 2, 3$ 。

解 $\forall a, b, 0 < a < b$, 有:

$$\begin{aligned} P(\mu - a < X < \mu + b) &= \int_{\mu-a}^{\mu+b} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \\ &= \int_{-\frac{a}{\sigma}}^{\frac{b}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \end{aligned}$$

这样在区间 $(\mu - k\sigma, \mu + k\sigma)$ 的概率 ($k = 1, 2, 3$) 为:

$$P(\mu - \sigma < X < \mu + \sigma) = \Phi(1) - \Phi(-1) = 0.682\ 6$$

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = \Phi(2) - \Phi(-2) = 0.954\ 4$$

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = \Phi(3) - \Phi(-3) = 0.997\ 4$$

其中, $\Phi(-x) = 1 - \Phi(x)$ 。从以上计算可知, 服从正态分布的随机变量 X 之值, 几乎都落在 $(\mu - \sigma, \mu + \sigma)$ 之间, 落在 $(\mu - 2\sigma, \mu + 2\sigma)$ 的机会较多。

例 1.2 某地水体 (C0) 农麦 $X \sim N(10, 4)$, 求 (C0) 农麦落在区间 $(8, 12)$ 的

例 1.1.2 查附表 1 求自由度为 199 的 χ^2 分布的上侧 α 分位点的近似值。

例如, 查附表 1 并计算, 可得:

$$\chi_{0.05}^2(199) \approx \frac{1}{2}(2.326\ 348 + \sqrt{199})^2 \approx 135.023\ 1$$

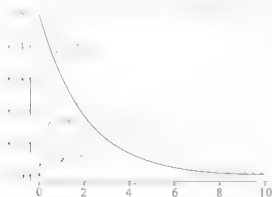


图 1.3 自由度分别为 2, 3, 5 的 χ^2 分布密度曲线图

1.1.3 t 分布

设 $X \sim N(0, 1)$, $Y \sim \chi^2(n)$, 并且 X, Y 独立, 则随机变量

$$T = \frac{X}{\sqrt{Y/n}}$$

服从自由度为 n 的 t 分布, 记为 $T \sim t(n)$ 。

对于任意一个 $t(n)$ 分布, 它的概率密度函数为:

$$P(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} \quad (-\infty < x < +\infty) \quad (1.6)$$

式中, n 为正整数。

可以证明, 当 $n \rightarrow \infty$ 时, $t(n)$ 分布趋近于标准正态分布 $N(0, 1)$ 。因此, 当自由度 n 较大时, $t(n)$ 分布与标准正态分布 $N(0, 1)$ 非常接近。在工程应用中, 当 $n \geq 30$ 时, 通常认为 $t(n)$ 分布与标准正态分布 $N(0, 1)$ 近似相等。

图 1.4 给出了不同自由度 n 下的 t 分布密度曲线。可以看出, 随着 n 的增加, 曲线越来越接近标准正态分布 $N(0, 1)$ 。

由式(1.7.1)和式(1.7.2)可得, t 分布的概率密度函数为

$$f_0(n_1, n_2) = \frac{1}{F_{1-\alpha}(n_2, n_1)} \quad (1.8)$$

由式(1.8)可得, t 分布的概率密度函数为 $f_0(n_1, n_2)$ 。由式(1.8)可得, t 分布的概率密度函数为 $F_{1-\alpha}(n_2, n_1)$ 。

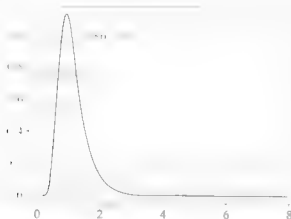


图 1.5 自由度为 3, 20 和 50, 20 的 t 分布密度曲线图

1.2 随机向量的数字特征

在概率论中, 随机向量的数字特征是描述随机向量的重要特征。本章首先对概率论中随机向量的主要数字特征作一介绍。

1.2.1 数学期望

“数学期望”或“期望”是随机变量的一个重要特征。它表示随机变量取值的平均值。在概率论中, 数学期望是描述随机变量取值的重要特征。本章首先对概率论中随机变量的主要数字特征作一介绍。

1.2.1.1 离散随机变量的数学期望

设 X 是一个离散随机变量, 其取值为 x_1, x_2, \dots, x_n , 对应的概率为 p_1, p_2, \dots, p_n 。

于是, 可得 X 的概率分布律为 p_1, p_2, \dots, p_k , 且 $\sum_{k=1}^{\infty} p_k = 1$, 故 X 的数学期望为

$$E(X) = \sum_{k=1}^{\infty} x_k p_k, \quad (k=1, 2, \dots)$$

$$E(X) = \sum_{k=1}^{\infty} x_k p_k, \quad (k=1, 2, \dots) \quad (1.9)$$

例 1.3 设 X 和 Y 分别表示甲、乙两条河流污染物的含量, 其概率分布律如表 1.1 所示, 试问哪条河流污染更重?

表 1.1 甲、乙两条河流 DO 的概率分布表

河流甲				河流乙					
X	0	1	2	3	Y	0	1	2	3
P	0.3	0.3	0.2	0.2	P	0.3	0.5	0.2	0.0

解

$$E(X) = 0 \times 0.3 + 1 \times 0.3 + 2 \times 0.2 + 3 \times 0.2 = 1.3$$

$$E(Y) = 0 \times 0.3 + 1 \times 0.5 + 2 \times 0.2 + 3 \times 0.0 = 0.9$$

上面结果表明, 河流乙污染更重。

1.2.1.2 连续随机变量的数学期望

设 X 为连续型随机变量, 其概率密度函数为 $f(x)$, 且 $f(x) \geq 0$, $\int_{-\infty}^{+\infty} f(x) dx = 1$, 则称 X 的数学期望存在, 且

$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx \quad (1.10)$$

例 1.4 设 X 为服从正态分布 $N(0, 1)$ 的随机变量, 求 $E(X)$ 。

求 $E(X)$ 。

解
$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx = \int_{-\infty}^{+\infty} x \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

因为 奇函数在对称区间的积分为 0

所以 $E(X) = 0$

1.2.1.3 随机变量函数的数学期望

设 X 为连续型随机变量, 其概率密度函数为 $f(x)$, $Y = g(X)$, 则 Y 的数学期望为

若 $(x_k) | p_k$ 收敛, 则 Y 的数学期望为:

$$E(Y) = E[g(X)] = \sum_{k=1}^{\infty} g(x_k) p_k \quad (1.11)$$

若 X 为连续型随机变量, 其概率密度函数为 $f(x)$, 则 $Y = g(X)$ 的数学期望定义为:

$$E(Y) = E[g(X)] = \int_{-\infty}^{+\infty} g(x) f(x) dx \quad (1.12)$$

例 1.5 设 X 为离散型随机变量, 其概率分布表如下, 求 $E(Y_1)$ 和 $E(Y_2)$ 。

表 1.2 随机变量 X 的概率分布表

X	1	-2	-1	0	1
P	0.1	0.3	0.4	0.2	

解 由表 1.2 可知 X 的概率分布为: $P\{X=1\}=0.1, P\{X=-2\}=0.3, P\{X=-1\}=0.4, P\{X=0\}=0.2$ 。

表 1.3 随机变量 Y_1 和 Y_2 的概率分布表

Y_1	-3	-1	1	3
P	0.1	0.3	0.4	0.2
Y_2	0	1	4	
P	0.4	0.5	0.1	

由公式(1.9)有:

$$E(Y_1) = (-3) \times 0.1 + (-1) \times 0.3 + 1 \times 0.4 + 3 \times 0.2 = 0.4$$

方法二:

方法二: 直接由公式(1.11)求 $E(Y_1)$, $E(Y_2)$ 。

$$E(Y_1) = E(2X+1) = [2 \times (-2) + 1] \times 0.1 + [2 \times (-1) + 1] \times 0.3 +$$

$$[2 \times 1 + 1] \times 0.4 +$$

1.2.1.4 性质

数学期望具有以下几个性质(a, b, c 均为常数):

(1) $E(c) = c$;

$$P \setminus \setminus I \setminus \setminus$$

$$(3) E(aX+b) = aE(X)+b.$$

1.2.2 方差和均方差

1.2.2.1 方差和均方差的定义

设 X 为离散型随机变量, 其可能取值为 x_1, x_2, \dots, x_n , 在 x_i 处出现的概率为 p_i , $i=1, 2, \dots, n$. 则 X 的方差定义为:

$$D(X) = E[X - E(X)]^2 \quad (1.13)$$

均方差定义为:

$$\sqrt{D(X)}$$

方差和均方差是衡量随机变量取值分散程度的.

1.2.2.2 离散型随机变量的方差

设 X 为离散型随机变量, 其可能取值为 x_1, x_2, \dots, x_n , 在 x_i 处出现的概率为 p_i , $i=1, 2, \dots, n$. 则 X 的方差表达式为:

$$D(X) = \sum [x_i - E(X)]^2 p_i \quad (1.14)$$

例 1.6 计算例 1.3 中的河流水质指数溶解氧的方差。

解 由上:

$$E(X^2) = 0^2 \times 0.3 + 1^2 \times 0.3 + 2^2 \times 0.2 + 3^2 \times 0.2 = 2.9$$

$$D(X) = E(X^2) - [E(X)]^2 = 2.9 - 1.3^2 = 1.21$$

$$E(Y^2) = 0^2 \times 0.3 + 1^2 \times 0.5 + 2^2 \times 0.2 + 3^2 \times 0.0 = 1.3$$

$$D(Y) = E(Y^2) - [E(Y)]^2 = 1.3 - 0.9^2 = 0.49$$

1.2.2.3 连续型随机变量的方差

设 X 为连续型随机变量, 其概率密度函数为 $f(x)$, 则 X 的方差表达式为:

$$D(X) = \int_{-\infty}^{+\infty} [x - E(X)]^2 P(x) dx \quad (1.15)$$

例 1.7 设随机变量 X 的概率密度为 $f(x)$, 且 X 与 X^2 相互独立, 求 $D(X)$ 。

$$P(x) = \begin{cases} \lambda e^{-\lambda x} & (x \geq 0) \\ 0 & (x < 0) \end{cases}$$

求 $D(X)$ 。

解

$$E(X) = \int_{-\infty}^{+\infty} x P(x) dx = \int_0^{+\infty} x \lambda e^{-\lambda x} dx = \frac{1}{\lambda}$$

$$D(X) = E(X^2) - [E(X)]^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$$

1.2.2.4 性质

方差具有以下几个性质(a, b, c 均为常数):

$$D(a) = 0;$$

$$D(X) = D(-X);$$

$$(3) D(aX+b) = a^2 D(X);$$

$$(4) D(X) = 0 \text{ 的充要条件是存在常数 } c, \text{ 使得 } P(X=c) = 1.$$

证(1)、(2) 由(1) 知 $P(a) = 1$, 故 $D(a) = D(X) = D(-X)$ 。

1.2.3 原点矩和中心矩

对于随机变量 X , 称 $E(X^k)$ 为 X 的 k 阶原点矩, $E[(X-a)^k]$ 为 X 的 k 阶中心矩, 其中 $a = E(X)$ 。特别地, $E(X)$ 为 X 的一阶原点矩, $E[(X-a)^2]$ 为 X 的二阶中心矩, 即方差。由此可知, 方差是描述随机变量取值分散程度的一个量。如果记数学期望为 $E(X)$, 则方差为:

$$D(X) = E(X^2) - [E(X)]^2$$

证: 由(1.15) 知 $D(X) = E[(X-E(X))^2] = E(X^2 - 2XE(X) + E(X)^2)$ 。

$$E(X) = \sum_{i=1}^n x_i p_i; E(X) = \int_{-\infty}^{+\infty} x f(x) dx \quad (1.16)$$

$$E(X^2) = \sum_{i=1}^n x_i^2 p_i; E(X^2) = \int_{-\infty}^{+\infty} x^2 f(x) dx \quad (1.17)$$

这样, 方差 $D(X)$ 的物理意义为: 随机变量 X 的取值与期望值的偏离程度。借用物理学中“质心”名称, $D(X)$ 为 $D(X) = E(X^2) - [E(X)]^2$, 故 $D(X)$ 为 X 的“质心矩”, 即 X 的

$$\Lambda(X) = \begin{pmatrix} \text{Cov}(X, Y_1) & \text{Cov}(X, Y_2) & \cdots & \text{Cov}(X, Y_p) \\ \text{Cov}(X_1, Y) & \text{Cov}(X_2, Y) & \cdots & \text{Cov}(X_n, Y) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_n, Y) & \text{Cov}(X_n, Y) & \cdots & \text{Cov}(X_n, Y) \end{pmatrix}$$

如果 $\text{Cov}(X, Y) = 0$, 则称 X 和 Y 是不相关的。

同样, 如果 $X = (X_1, X_2, \dots, X_n)^T$ 为随机向量, 那么 $\Lambda(X)$ 就相应转化为如下形式:

$$\Lambda(X) = \begin{pmatrix} D(X_1) & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & D(X_2) & \cdots & \text{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \text{Cov}(X_n, X_2) & \cdots & D(X_n) \end{pmatrix} \quad (1.22)$$

称上述矩阵为随机向量 X 的自协方差阵。

1.2.6 随机变量的相关系数

设 $(X, Y) = (X(X), Y(Y))$ 为二维随机变量, 若存在 $D(X) > 0, D(Y) > 0$, 则相关系数 ρ_{xy} 定义为:

$$\rho_{xy} = \frac{\text{Cov}(X, Y)}{\sqrt{D(X)}\sqrt{D(Y)}} \quad (1.23)$$

相关系数描述了随机变量之间的相关程度。

当 $\rho_{xy} = 1$ 时, 称 X 和 Y 完全正相关; 当 $\rho_{xy} = -1$ 时, 称 X 和 Y 完全负相关; 当 $\rho_{xy} = 0$ 时, 称 X 和 Y 完全不相关。如果 X 和 Y 完全正相关, 则存在常数 a 和 b , 使得 $Y = aX + b$ 。

当 ρ_{xy} 介于 -1 和 1 之间时, 称 X 和 Y 不完全相关。当 ρ_{xy} 为正时, 称 X 和 Y 正相关; 当 ρ_{xy} 为负时, 称 X 和 Y 负相关。当 $\rho_{xy} = 0$ 时, 称 X 和 Y 不相关。

例 1.8 设 X 和 Y 为二维随机变量, $X \sim N(0, 1)$, $Y \sim N(0, 1)$, 且 $\text{Cov}(X, Y) = 0$, 则 X 和 Y 是否独立? 解: 因为 X 和 Y 都是标准正态分布, 且 $\text{Cov}(X, Y) = 0$, 所以 X 和 Y 是独立的。

表 1.4 某河流水质监测结果 单位: mg/L

点号	1#	2#	3#	4#
1#	2	3	5	8
2#	3	5	8	8
3#	19	19	4	4
4#	3	6	7	7

解 根据协方差定义, 计算得表 1.5。

表 1.5 协方差阵元素所需数据表

m	X_i	X_j	X_k	$X_i - \bar{X}_i$	$X_j - \bar{X}_j$	$X_k - \bar{X}_k$
1	2	3	8	-4.75	-5.25	1.25
2	3	5	8	3.75	-3.25	1.25
3	19	19	4	12.25	10.75	-2.75
4	3	6	7	-3.75	-2.25	0.25
\bar{X}	6.75	8.25	6.75			

$$\sigma_{X_i, X_i} = E\{[X_i - E(X_i)]^2\} = \frac{1}{n} \sum_{m=1}^n (X_{im} - \bar{X}_i)^2 = 44.3125$$

$$\sigma_{X_i, X_j} = E\{[X_i - E(X_i)][X_j - E(X_j)]\} = \frac{1}{n} \sum_{m=1}^n (X_{im} - \bar{X}_i)(X_{jm} - \bar{X}_j) = -11.3125$$

$$\sigma_{X_i, X_k} = E\{[X_i - E(X_i)][X_k - E(X_k)]\} = \frac{1}{n} \sum_{m=1}^n (X_{im} - \bar{X}_i)(X_{km} - \bar{X}_k) = -10.1875$$

$$\sigma_{X_i, X_i} = E\{[X_i - E(X_i)]^2\} = \frac{1}{n} \sum_{m=1}^n (X_{im} - \bar{X}_i)^2 = 50.1875$$

$$\sigma_{X_j, X_j} = E\{[X_j - E(X_j)]^2\} = \frac{1}{n} \sum_{m=1}^n (X_{jm} - \bar{X}_j)^2 = 39.6875$$

$$\sigma_{X_k, X_k} = E\{[X_k - E(X_k)]^2\} = \frac{1}{n} \sum_{m=1}^n (X_{km} - \bar{X}_k)^2 = 2.6875$$

即协方差阵为:

$$V = \begin{bmatrix} 50.1875 & 44.3125 & 11.3125 \\ 44.3125 & 39.6875 & -10.1875 \\ -11.3125 & -10.1875 & 2.6875 \end{bmatrix}$$

根据相关系数的定义, 得相关系数矩阵为:

$$R = \begin{bmatrix} 1.000 & 0.993 & -0.974 \\ 0.993 & 1.000 & 0.986 \\ -0.974 & -0.986 & 1.000 \end{bmatrix}$$

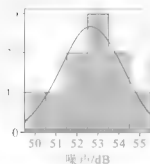


图 1-6 频率直方图

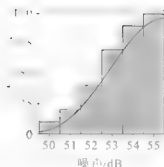


图 1-7 累积频率图

近似曲线(图 1-6~1-7)。

1.2.10 中心极限定理

设总体 X 服从正态分布 $N(\mu, \sigma^2)$, 则 X 的样本均值 \bar{X} 服从正态分布 $N(\mu, \sigma^2/n)$, 且 \bar{X} 的分布密度函数为 $f(\bar{x}) = \frac{1}{\sigma\sqrt{n}} \exp\left\{-\frac{(\bar{x}-\mu)^2}{2\sigma^2/n}\right\}$ 。中心极限定理(康永尚等, 2005)。

例 1.10 设 X_1, X_2, \dots, X_n 为独立同分布的随机变量, 且 $X_i \sim N(\mu, \sigma^2)$, 求 \bar{X} 的分布密度函数。

表 1.7 随机变量 X (总体)的取值、期望值和方差

序号	X 的取值	X 的期望值	X 的方差
1	332	$\mu = E(X) = \frac{1}{5} \sum_{i=1}^5 x_i = 340$	$D(X) = \sigma^2 = \frac{1}{5} \sum_{i=1}^5 (x_i - 340)^2 = 32$
2	336		
3	340		
4	344		
5	348		

解 由表 1.7 可知, X 的期望值和方差分别为 $\mu = 340$ 和 $\sigma^2 = 32$, 故 \bar{X} 的分布密度函数为 $f(\bar{x}) = \frac{1}{\sigma\sqrt{n}} \exp\left\{-\frac{(\bar{x}-\mu)^2}{2\sigma^2/n}\right\}$ 。组合、均值及均值统计量列于表 1.8 中。

表 1.8 样本点组合、变量值组合、均值及均值统计量

样本序号	样本点组合	变量值组合	样本均值	样本均值统计量
1	(1, 1)	(332, 332)	332	$E(\bar{X}) = \frac{1}{25} \sum_{i=1}^{25} \bar{x}_i = 340$
2	(1, 2)	(332, 336)	334	
3	(1, 3)	(332, 340)	336	
4	(1, 4)	(332, 344)	338	
5	(1, 5)	(332, 348)	340	$D(\bar{X}) = \frac{1}{25} \sum_{i=1}^{25} [\bar{x}_i - E(\bar{X})]^2 = 16$
6	(2, 1)	(336, 332)	334	
7	(2, 2)	(336, 336)	336	
8	(2, 3)	(336, 340)	338	
9	(2, 4)	(336, 344)	340	
10	(2, 5)	(336, 348)	342	
11	(3, 1)	(340, 332)	336	
12	(3, 2)	(340, 336)	338	
13	(3, 3)	(340, 340)	340	
14	(3, 4)	(340, 344)	342	
15	(3, 5)	(340, 348)	344	
16	(4, 1)	(344, 332)	338	
17	(4, 2)	(344, 336)	340	
18	(4, 3)	(344, 340)	342	
19	(4, 4)	(344, 344)	344	
20	(4, 5)	(344, 348)	346	
21	(5, 1)	(348, 332)	340	
22	(5, 2)	(348, 336)	342	
23	(5, 3)	(348, 340)	344	
24	(5, 4)	(348, 344)	346	
25	(5, 5)	(348, 348)	348	

表 1.8 中, 样本点组合、变量值组合、样本均值、样本均值统计量, 分别对应于式 (1.1) 中的 ω 、 X 、 \bar{x} 、 \bar{X} 。由表 1.8 可知, 样本点组合共有 25 种, 变量值组合共有 25 种, 样本均值共有 25 种, 样本均值统计量共有 25 种。

例 1.3.1 设总体 X 服从正态分布 $N(\mu, \sigma^2)$, 总体均值 μ 的未知, 方差 σ^2 未知, 求 μ 的极大似然估计值。

解 设 x_1, x_2, \dots, x_n 为 X 的 n 个独立样本, 把数代入式 (1.3.1) 有

$$L(\mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sigma \sqrt{2\pi}} \exp\left\{-\frac{(x_i - \mu)^2}{2\sigma^2}\right\} \quad (1.3.2)$$

个估计值。

对式 (1.3.2) 求对数似然函数, 得

$$\ln L(\mu, \sigma^2) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \quad (1.3.3)$$

对式 (1.3.3) 求偏导数, 令

$$\frac{\partial \ln L(\mu, \sigma^2)}{\partial \mu} = 0, \quad \frac{\partial \ln L(\mu, \sigma^2)}{\partial \sigma^2} = 0 \quad (1.3.4)$$

解得

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2 \quad (1.3.5)$$

式 (1.3.5) 即为 μ 和 σ^2 的极大似然估计值。由此可知, 极大似然估计法可估计总体均值 (μ)、总体标准差 (σ)。

本书中只对参数估计部分作简单的介绍。

1.3.1 点估计

设 X_1, X_2, \dots, X_n 为独立同分布的随机变量, 且 X_i 服从正态分布 $N(\mu, \sigma^2)$, 其中 μ 未知, σ^2 已知, 求 μ 的极大似然估计值。解 设 x_1, x_2, \dots, x_n 为 X_1, X_2, \dots, X_n 的 n 个独立样本, 代入式 (1.3.1) 有

$$L(\mu) = \prod_{i=1}^n \frac{1}{\sigma \sqrt{2\pi}} \exp\left\{-\frac{(x_i - \mu)^2}{2\sigma^2}\right\} \quad (1.3.6)$$

对式 (1.3.6) 求对数似然函数, 得

$$\ln L(\mu) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \quad (1.3.7)$$

对式 (1.3.7) 求偏导数, 令

$$\frac{\partial \ln L(\mu)}{\partial \mu} = 0 \quad (1.3.8)$$

解得

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i \quad (1.3.9)$$

式 (1.3.9) 即为 μ 的极大似然估计值。由此可知, 极大似然估计法可估计总体均值 (μ)、总体标准差 (σ)。

1.3.2 区间估计

设 X_1, X_2, \dots, X_n 为独立同分布的随机变量, 且 X_i 服从正态分布 $N(\mu, \sigma^2)$, 其中 μ 未知, σ^2 已知, 求 μ 的极大似然估计值。

解 设 x_1, x_2, \dots, x_n 为 X_1, X_2, \dots, X_n 的 n 个独立样本, 代入式 (1.3.1) 有

$$L(\mu) = \prod_{i=1}^n \frac{1}{\sigma \sqrt{2\pi}} \exp\left\{-\frac{(x_i - \mu)^2}{2\sigma^2}\right\} \quad (1.3.10)$$

对式 (1.3.10) 求对数似然函数, 得

$$\ln L(\mu) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \quad (1.3.11)$$

对式 (1.3.11) 求偏导数, 令

$$\frac{\partial \ln L(\mu)}{\partial \mu} = 0 \quad (1.3.12)$$

解得

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i \quad (1.3.13)$$

且 $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ 为样本 X_1, X_2, \cdots, X_n 的次序统计量, 即第 i 个最小值, 称为第 i 个次序统计量, 为区间 $(-\infty, x_{(i)}]$ 的估计量, 记为 $\hat{F}_n(x) = x_{(i)}/n$, 其中 $i = 1, 2, \cdots, n$ 。

由式 (1.24) 可得 $1 - \alpha = P\left\{-x_{(\alpha)} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq x_{(1-\alpha)}\right\}$, 又由式 (1.23) 可得 $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} + \frac{\sigma}{\sqrt{n}} \frac{x_{(\alpha)}}{\sigma} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} + \frac{\sigma}{\sqrt{n}} \frac{x_{(\alpha)}}{\sigma}$ 。

$$P\left\{-x_{(\alpha)} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq x_{(1-\alpha)}\right\} = 1 - \alpha \quad (1.25)$$

上式经过变换后可得:

$$P\left\{\bar{X} - \frac{\sigma}{\sqrt{n}} x_{(\alpha)} \leq \mu \leq \bar{X} + \frac{\sigma}{\sqrt{n}} x_{(1-\alpha)}\right\} = 1 - \alpha \quad (1.26)$$

上式即表示: 在 $(1 - \alpha)$ 置信水平下总体期望值的区间估计为:

$$\bar{X} \pm \frac{\sigma}{\sqrt{n}} x_{(\alpha)} \leq \mu \leq \bar{X} + \frac{\sigma}{\sqrt{n}} x_{(1-\alpha)} \quad (1.27)$$

例 1.11 设 X_1, X_2, \cdots, X_n 为独立同分布的随机变量, 且 $X_i \sim N(\mu, \sigma^2)$, 其中 μ, σ^2 未知。

已知 $n = 5, \bar{X} = 5, x_{(0.025)} = 1.96$, 求参数 μ 的区间估计。

解 由式 (1.26) 可得

$$\bar{X} \pm \frac{\sigma}{\sqrt{n}} x_{(\alpha)} = 5 \pm \frac{\sigma}{\sqrt{5}} \times 1.96$$

$$\bar{X} \pm \frac{\sigma}{\sqrt{n}} x_{(\alpha)} = 5 \pm \frac{\sigma}{\sqrt{5}} \times 1.96$$

因为:

$$\bar{X} = 5, n = 5, x_{(0.025)} = 1.96$$

所以, 参数 μ 的区间估计为 $5 \pm \frac{\sigma}{\sqrt{5}} \times 1.96$ 。

在估计量中, 估计量 $\hat{\theta}$ 的无偏性是指 $E(\hat{\theta}) = \theta$, 即估计量的期望值等于被估计参数的真值。下面就简单介绍一下估计量的几个评价标准。

1.3.2.1 无偏估计量

设 $\hat{\theta} = \hat{\theta}(X_1, X_2, \cdots, X_n)$ 是未知参数 θ 的一个估计量, 若

$$E(\hat{\theta}) = \theta$$

则称 $\hat{\theta}$ 为 θ 的无偏估计量。

在估计量中, 估计量 $\hat{\theta}$ 的无偏性是指 $E(\hat{\theta}) = \theta$, 即估计量的期望值等于被估计参数的真值。下面就简单介绍一下估计量的几个评价标准。

又因为无偏估计量具有唯一性, 所以 \bar{X} 是 μ 的无偏估计量。同理可证 S^2 也是 σ^2 的无偏估计量。因此, \bar{X} 和 S^2 都是 μ 和 σ^2 的无偏估计量, 且 \bar{X} 和 S^2 的期望值基本上相同。

当一个估计量不是无偏估计量时, 称它为有偏估计量。

下面来讨论估计量的有效性。有效性是指估计量的方差, 方差越小, 估计量越有效。

因为 $X = \frac{1}{n} \sum_{i=1}^n X_i$, 所以 $E(X) = \mu$, $D(X) = \frac{1}{n^2} \sum_{i=1}^n D(X_i) = \frac{\sigma^2}{n}$ 。从而

$$E(X) = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \mu$$

所以 X 是 μ 的无偏估计量。

$$D(X) = D\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n D(X_i) = \frac{\sigma^2}{n}$$

$$E(\bar{X}^2) = D(\bar{X}) + E(\bar{X})^2 = \frac{1}{n} \sigma^2 + \mu^2$$

所以 $E(S^2) = \frac{n-1}{n} \sigma^2$

因此, S^2 也是 σ^2 的无偏估计量。当 n 很大时, $\frac{n-1}{n} \sigma^2 \approx \sigma^2$, 所以 S^2 也是 σ^2 的统计量称之为渐近统计量。

下面来讨论估计量的有效性。有效性是指估计量的方差, 方差越小, 估计量越有效。

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

则此时

$$D(S^2) = \frac{1}{(n-1)^2} D\left(\sum_{i=1}^n (X_i - \bar{X})^2\right)$$

由 $D(X) = \frac{\sigma^2}{n}$ 可知 $D(X_i) = \frac{\sigma^2}{n}$, 所以 $D(X_i - \bar{X}) = \frac{\sigma^2}{n}$ 。

因此, S^2 是 σ^2 的无偏估计量, 且 S^2 的方差为 $\frac{2\sigma^4}{n-1}$, 所以 S^2 是 σ^2 的渐近统计量。

1.3.2.2 一致估计

设 X_1, X_2, \dots, X_n 是独立同分布的随机变量, μ 和 σ^2 是 X_i 的期望和方差, 且 μ 和 σ^2 是 X_i 的无偏估计量, 且 μ 和 σ^2 是 X_i 的渐近统计量, 则称 μ 和 σ^2 为

找出假设与现实之间的矛盾,从而否定这个假设。

1.4.1 假设检验的原理

假设检验是统计推断的重要内容之一,其基本思想是“反证法”,即先假设原假设 H_0 为真,然后根据样本信息,检验原假设是否成立。如果原假设成立,则接受原假设;如果原假设不成立,则拒绝原假设,接受备择假设 H_1 。假设检验的原理可以概括为以下几点:

1. 提出假设:根据问题的性质,提出原假设 H_0 和备择假设 H_1 。

2. 构造检验统计量:根据样本数据,构造一个能够反映原假设是否成立的统计量。

3. 确定拒绝域:根据检验统计量的分布,确定一个临界值,使得当检验统计量的值大于临界值时,拒绝原假设。

4. 做出决策:根据样本数据,计算检验统计量的值,并与临界值进行比较,做出接受或拒绝原假设的决策。

假设检验的原理可以用一个例子来说明。假设有一个骗子,他声称自己从来没有骗过人。为了验证他的说法,我们设计了一个假设检验。原假设 H_0 是“骗子从来没有骗过人”,备择假设 H_1 是“骗子曾经骗过人”。我们收集了一些数据,发现骗子在 10 次中骗了 3 次。根据这些数据,我们构造了一个检验统计量,即骗子的次数。如果骗子的次数大于 3,我们就拒绝原假设,认为骗子曾经骗过人。如果骗子的次数小于等于 3,我们就接受原假设,认为骗子从来没有骗过人。

假设检验的原理可以用一个数学模型来表示。假设有一个随机变量 X ,其分布函数为 $F(x)$ 。我们提出原假设 $H_0: F(x) = F_0(x)$ 和备择假设 $H_1: F(x) \neq F_0(x)$ 。我们构造一个检验统计量 T ,其分布函数为 $G(t)$ 。我们确定一个临界值 c ,使得当 $T > c$ 时,拒绝原假设。假设检验的原理可以表示为:

有 $H_0: F(x) = F_0(x)$ 和 $H_1: F(x) \neq F_0(x)$ 。

构造检验统计量 T ,其分布函数为 $G(t)$ 。

确定临界值 c ,使得当 $T > c$ 时,拒绝原假设。

做出决策:如果 $T > c$,则拒绝原假设;否则,接受原假设。

假设检验的原理可以用一个数学公式来表示。假设有一个随机变量 X ,其分布函数为 $F(x)$ 。我们提出原假设 $H_0: F(x) = F_0(x)$ 和备择假设 $H_1: F(x) \neq F_0(x)$ 。我们构造一个检验统计量 T ,其分布函数为 $G(t)$ 。我们确定一个临界值 c ,使得当 $T > c$ 时,拒绝原假设。假设检验的原理可以表示为:

有 $H_0: F(x) = F_0(x)$ 和 $H_1: F(x) \neq F_0(x)$ 。

构造检验统计量 T ,其分布函数为 $G(t)$ 。

确定临界值 c ,使得当 $T > c$ 时,拒绝原假设。

做出决策:如果 $T > c$,则拒绝原假设;否则,接受原假设。

$$P(\text{拒绝 } H_0 | H_0 \text{ 为真}) \leq \alpha$$

在假设检验中,我们通常设定一个显著性水平 α ,即当原假设为真时,拒绝原假设的概率不超过 α 。这个概率称为第一类错误的概率,通常记为 α 。即:

概率称为第二类错误的概率,通常记为 β ,即:

$$P(\text{接受 } H_0 | H_0 \text{ 不真})$$

假设检验的原理可以用一个数学公式来表示。假设有一个随机变量 X ,其分布函数为 $F(x)$ 。我们提出原假设 $H_0: F(x) = F_0(x)$ 和备择假设 $H_1: F(x) \neq F_0(x)$ 。我们构造一个检验统计量 T ,其分布函数为 $G(t)$ 。我们确定一个临界值 c ,使得当 $T > c$ 时,拒绝原假设。假设检验的原理可以表示为:

1.4.3 参数检验

参数检验是指对总体的某些参数(如均值、方差、比例等)是否等于某个已知值,或是否等于某个已知值之间的差异进行检验,并作出统计推断。

1.4.3.1 总体方差 σ^2 已知, 检验总体均值 μ

设总体 $X \sim N(\mu, \sigma^2)$, σ^2 已知, 从总体 X 中抽取样本 X_1, X_2, \dots, X_n , 样本均值 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, 检验 $H_0: \mu = \mu_0$ 和 $H_1: \mu \neq \mu_0$ 。

$$(1) H_0: \mu = \mu_0; H_1: \mu \neq \mu_0$$

$$(2) H_0: \mu \geq \mu_0; H_1: \mu < \mu_0$$

$$(3) H_0: \mu \leq \mu_0; H_1: \mu > \mu_0$$

其中, μ_0 为已知值, μ_1 为待检验值, μ_2 为待检验值, μ_3 为待检验值, μ_4 为待检验值, μ_5 为待检验值, μ_6 为待检验值, μ_7 为待检验值, μ_8 为待检验值, μ_9 为待检验值。

检验统计量为 $Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$, 检验统计量 Z 服从标准正态分布 $N(0, 1)$, 检验统计量 Z 的分布函数为 $\Phi(z)$, 检验统计量 Z 的临界值为 $z_{\alpha/2}$, 检验统计量 Z 的拒绝域为 $|Z| \geq z_{\alpha/2}$, 检验统计量 Z 的接受域为 $|Z| < z_{\alpha/2}$ 。

表 1.9 σ^2 已知时单个总体均值的 Z 检验法

检验方法	双侧检验	单侧检验	
原假设 H_0	$\mu = \mu_0$	$\mu \geq \mu_0$	$\mu \leq \mu_0$
备择假设 H_1	$\mu \neq \mu_0$	$\mu < \mu_0$	$\mu > \mu_0$
检验统计量	$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$		
临界值 C	$z_{\alpha/2}$	z_{α}	z_{α}
拒绝域	$ Z \geq C$	$Z \leq C$	$Z \geq C$

例 1.12 为了比较两种不同浓度的氯化钠溶液对某种细菌的抑制作用, 分别抽取了 10 个样本, 测得细菌的抑制率, 数据如下表所示。问两种浓度的氯化钠溶液对细菌的抑制作用是否有显著差异?

解 由于 $\mu_0 = 250$ mg/L, $\sigma_0 = 10$ mg/L, 故又为方差已知且服从正态分布, 这是一个单侧的正态检验。

已知条件: $\mu_0 = 250$ mg/L, 显著性水平 $\alpha = 0.05$, $\sigma_0 = 10$, $n = 25$

根据数据, 计算得到: $\bar{X} = 279$ mg/L。

假设检验过程为:

$$H_0$$

$$H_1$$

检验统计量为:

$$\frac{\bar{X} - \mu_0}{\sigma_0 / \sqrt{n}} = \frac{279 - 250}{10 / \sqrt{25}}$$

查表得: $z_{1-\alpha} = z_{1-0.05} = z_{0.95} = 1.645$, 故统计量落在拒绝域, 即 H_0 不成立, 认为该厂污水中铜含量超标。

1.4.3.2 总体方差 σ 未知, 检验总体均值 μ

设总体 $X \sim N(\mu, \sigma^2)$, σ^2 未知, 从总体 X 中抽取容量为 n 的样本 X_1, X_2, \dots, X_n , 且 X_1, X_2, \dots, X_n 相互独立, 记 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ 为样本均值, $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ 为样本方差, 则有以下两种检验假设:

$$H_0: \mu = \mu_0; H_1: \mu \neq \mu_0$$

(2) $H_0: \mu \geq \mu_0; H_1: \mu < \mu_0$

(3) $H_0: \mu \leq \mu_0; H_1: \mu > \mu_0$

由于 σ^2 未知, 故用样本方差 S^2 代替总体方差 σ^2 , 此时统计量 $T = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$ 服从自由度为 $n-1$ 的 t 分布, 即 $T \sim t_{n-1}$ 。与 z 检验类似, 对于给定的显著性水平 α , 查表得 $t_{\alpha/2, n-1}$ 和 $t_{1-\alpha/2, n-1}$, 从而得到双侧检验的拒绝域为 $|T| > t_{\alpha/2, n-1}$ 。

对于单侧检验, 查表得 $t_{\alpha, n-1}$ 和 $t_{1-\alpha, n-1}$, 从而得到单侧检验的拒绝域为 $T > t_{\alpha, n-1}$ 或 $T < t_{1-\alpha, n-1}$ 。对于双侧检验, 查表得 $t_{\alpha/2, n-1}$ 和 $t_{1-\alpha/2, n-1}$, 从而得到双侧检验的拒绝域为 $|T| > t_{\alpha/2, n-1}$ 。以上即为总体均值的 t 检验法。

表 1.10 σ^2 未知时单个总体均值的 t 检验法

检验方法	双侧检验		单侧检验
原假设 H_0	$\mu = \mu_0$	$\mu \geq \mu_0$	$\mu \leq \mu_0$
备择假设 H_1	$\mu \neq \mu_0$	$\mu < \mu_0$	$\mu > \mu_0$
检验统计量	$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$		
临界值 C	$t_{\alpha/2} (n-1)$	$-t_{\alpha} (n-1)$	$t_{\alpha} (n-1)$
统计推断 拒绝域	$ t > C$	$t < C$	$t > C$

例 1.13 某厂生产的一种零件，按规定其长度应服从正态分布，且方差为 0.054。现从该厂生产的一批零件中随机抽取 5 件，测得其长度分别为 4.55, 4.68, 4.72, 4.78, 4.82。问该厂生产的零件长度是否符合规定？

解 设该厂生产的零件长度 X 服从正态分布 $N(\mu, \sigma^2)$ ，其中 $\sigma^2 = 0.054$ 。现从该厂生产的一批零件中随机抽取 5 件，测得其长度分别为 4.55, 4.68, 4.72, 4.78, 4.82。问该厂生产的零件长度是否符合规定？

解 设该厂生产的零件长度 X 服从正态分布 $N(\mu, \sigma^2)$ ，其中 $\sigma^2 = 0.054$ 。现从该厂生产的一批零件中随机抽取 5 件，测得其长度分别为 4.55, 4.68, 4.72, 4.78, 4.82。问该厂生产的零件长度是否符合规定？

有关已知条件为： $\mu_0 = 4.55$ ， $n = 5$ ， $\bar{X} = 4.364$ ， $S = 0.054$

则检测过程为：

$$H_0: \mu = \mu_0 = 4.55; \quad H_1: \mu \neq \mu_0$$

检验统计量为：

$$t = \frac{|\bar{X} - \mu_0|}{S/\sqrt{n}} = \frac{|4.364 - 4.55|}{0.054/\sqrt{5}} = 7.702$$

因为 $t = 7.702 > t_{\alpha/2} (n-1) = t_{0.05/2} (5-1) = 2.776$ ，所以拒绝 H_0 ，认为该厂生产的零件长度不符合规定。

1.4.3.3 一个正态总体方差的假设检验

设总体 X 服从正态分布 $N(\mu, \sigma^2)$ ，其中 μ 未知， σ^2 未知。现从该总体 X 中随机抽取 n 个样本 X_1, X_2, \dots, X_n ，测得其方差分别为 $s_1^2, s_2^2, \dots, s_n^2$ 。问该总体 X 的方差是否符合规定？

$$(1) H_0: \sigma^2 = \sigma_0^2; \quad H_1: \sigma^2 \neq \sigma_0^2$$

$$(2) H_0: \sigma^2 \geq \sigma_0^2; \quad H_1: \sigma^2 < \sigma_0^2$$

$$(3) H_0: \sigma^2 \leq \sigma_0^2; \quad H_1: \sigma^2 > \sigma_0^2$$

解 设该总体 X 服从正态分布 $N(\mu, \sigma^2)$ ，其中 μ 未知， σ^2 未知。现从该总体 X 中随机抽取 n 个样本 X_1, X_2, \dots, X_n ，测得其方差分别为 $s_1^2, s_2^2, \dots, s_n^2$ 。问该总体 X 的方差是否符合规定？

1.4.3.4 两个正态总体参数的假设检验

设两个独立正态总体 $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, 分别来自两个总体抽取容量为 n_1, n_2 的样本, 样本均值 \bar{x}, \bar{y} , 样本方差 s_1^2, s_2^2 , 检验统计量 $T = \frac{\bar{x} - \bar{y} - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$, 检验统计量 T 服从自由度为 $n_1 + n_2 - 2$ 的 t 分布, 记为 $T \sim t(n_1 + n_2 - 2)$, 与方差的差异性分别予以讨论。

1. 两个正态总体均值差异性检验

设两个独立正态总体 $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, 分别来自两个总体抽取容量为 n_1, n_2 的样本, 检验统计量 T 服从自由度为 $n_1 + n_2 - 2$ 的 t 分布, 记为 $T \sim t(n_1 + n_2 - 2)$, 与方差的差异性分别予以讨论。

$$(1) H_0: \mu_1 = \mu_2; H_1: \mu_1 \neq \mu_2$$

$$(2) H_0: \mu_1 \geq \mu_2; H_1: \mu_1 < \mu_2$$

$$(3) H_0: \mu_1 \leq \mu_2; H_1: \mu_1 > \mu_2$$

检验统计量 $T = \frac{\bar{x} - \bar{y} - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$, 检验统计量 T 服从自由度为 $n_1 + n_2 - 2$ 的 t 分布, 记为 $T \sim t(n_1 + n_2 - 2)$, 与方差的差异性分别予以讨论。

表 1.13 两个独立总体均值比较的 Z 检验和 t 检验, α 水平

检验方法	Z 检验		t 检验	
适用情景	σ_1^2, σ_2^2 已知		σ_1^2, σ_2^2 未知	
原假设 H_0	$\mu_1 = \mu_2$	$\mu_1 \leq \mu_2$ 或 $\mu_1 \geq \mu_2$	$\mu_1 = \mu_2$	$\mu_1 \leq \mu_2$ 或 $\mu_1 \geq \mu_2$
备择假设 H	$\mu_1 \neq \mu_2$	$\mu_1 > \mu_2$ 或 $\mu_1 < \mu_2$	$\mu_1 \neq \mu_2$	$\mu_1 > \mu_2$ 或 $\mu_1 < \mu_2$
检验统计量	$Z = \frac{\bar{x} - \bar{y} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$		$T = \frac{\bar{x} - \bar{y} - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$	
检验临界值 C	$z_{\alpha/2}$	z_{α}	$t_{\alpha/2}(n_1 + n_2 - 2)$	$t_{\alpha}(n_1 + n_2 - 2)$

例 1.15 设两个独立正态总体 $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, 分别来自两个总体抽取容量为 n_1, n_2 的样本, 样本均值 \bar{x}, \bar{y} , 样本方差 s_1^2, s_2^2 , 检验统计量 $T = \frac{\bar{x} - \bar{y} - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$, 检验统计量 T 服从自由度为 $n_1 + n_2 - 2$ 的 t 分布, 记为 $T \sim t(n_1 + n_2 - 2)$, 与方差的差异性分别予以讨论。

解 由题意知, 两个独立正态总体 $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, 分别来自两个总体抽取容量为 n_1, n_2 的样本, 样本均值 \bar{x}, \bar{y} , 样本方差 s_1^2, s_2^2 , 检验统计量 $T = \frac{\bar{x} - \bar{y} - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$, 检验统计量 T 服从自由度为 $n_1 + n_2 - 2$ 的 t 分布, 记为 $T \sim t(n_1 + n_2 - 2)$, 与方差的差异性分别予以讨论。

有关已知条件为: $n_1 = 63$, $\bar{x}_1 = 62.3$, $\sigma_1^2 = 10.8$; $n_2 = 74$, $\bar{x}_2 = 66.8$,

则检测过程为:

$$H: \mu_1 = \mu_2; H_1: \mu_1 \neq \mu_2$$

检验统计量为:

$$Z = \frac{|x_1 - x_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{|62.3 - 66.8|}{\sqrt{\frac{1.5^2}{10} + \frac{1.5^2}{10}}} = 7.59$$

因为 $|Z| > Z_{\alpha/2}$, 所以拒绝 H_0 , 认为甲、乙两地放射污染强度显著不同。

2. 两个正态总体方差的差异性检验

如下:

$$(1) H_0: \sigma_1^2 = \sigma_2^2; H_1: \sigma_1^2 \neq \sigma_2^2$$

$$(2) H_0: \sigma_1^2 \leq \sigma_2^2; H_1: \sigma_1^2 > \sigma_2^2$$

$$(3) H_0: \sigma_1^2 \geq \sigma_2^2; H_1: \sigma_1^2 < \sigma_2^2$$

表 1.14 两个总体方差的 F 检验 (α 水平)

检验方法	双侧检验	单侧检验	
原假设 H_0	$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \leq \sigma_2^2$	$\sigma_1^2 \geq \sigma_2^2$
备择假设 H_1	$\sigma_1^2 \neq \sigma_2^2$	$\sigma_1^2 > \sigma_2^2$	$\sigma_1^2 < \sigma_2^2$
检验统计量	$F = S_1^2 / S_2^2$	$F = S_1^2 / S_2^2$	$F = S_1^2 / S_2^2$
n_1	$n_1 - 1$	$n_1 - 1$	$n_1 - 1$
n_2	$n_2 - 1$	$n_2 - 1$	$n_2 - 1$
检验临界值	$F_{\alpha/2}(n_1 - 1, n_2 - 1),$ $F_{1-\alpha/2}(n_1 - 1, n_2 - 1)$	$F_{\alpha}(n_1 - 1, n_2 - 1)$	$F_{1-\alpha}(n_1 - 1, n_2 - 1)$
拒绝域	$F > F_{\alpha/2}$ 或 $F < F_{1-\alpha/2}$	$F > F_{\alpha}$	$F < F_{1-\alpha}$

例 1.16 为比较两种脱硫装置的脱硫效率, 分别抽取脱硫装置 A、B 的脱硫效率数据, 经计算得到 $\bar{x}_A = 82.5$, $s_A^2 = 1.5$, $\bar{x}_B = 79.5$, $s_B^2 = 1.5$, $n_A = 10$, $n_B = 10$, 试比较两种脱硫装置的脱硫效率是否有显著差异。

表 1.15 脱硫装置的脱硫效率

型号 1	98	82	96
型号 2	92	95	89

解 $t = \frac{\bar{x} - \mu_0}{S/\sqrt{n}} = \frac{1.2 - 1}{8/\sqrt{10}} = 0.375$, 查表得 $t_{0.05}(9) = 1.833$, $t_{0.01}(9) = 2.821$, $n_2 = 3$, $S_2^2 = 9$.

检验过程为:

$$H_0: \sigma_1^2 = \sigma_2^2; H_1: \sigma_1^2 \neq \sigma_2^2$$

由于: $S_1^2 > S_2^2$, 则检验统计量

$$\frac{S_1^2}{S_2^2} = 76/9 = 8.44$$

查表得 $F_{0.05}(10, 2) = 4.10$, $F_{0.01}(10, 2) = 9.01$, 因为 $8.44 < 9.01$, 故接受 H_0 , 即认为两总体方差相等.

1.4.3.5 总结

以上介绍了正态总体参数的显著性假设检验, 下面将常用的假设检验方法总结如下表.

表 1.16 正态总体参数的显著性假设检验

检验参数	假设 H_1	统计量	分布
个 体 σ^2	$\mu = \mu_0$ (σ^2 已知)	$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	$N(0, 1)$
	$\mu = \mu_0$ (σ^2 未知)	$t = \frac{\bar{x} - \mu_0}{S/\sqrt{n}}$	$t(n-1)$
	$\sigma^2 = \sigma_0^2$ (μ 已知)	$\chi^2 = \frac{\sum_{i=1}^n (x_i - \mu_0)^2}{\sigma_0^2}$	$\chi^2(n)$
	$\sigma^2 = \sigma_0^2$ (μ 未知)	$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$	$\chi^2(n-1)$
两 总 体	$\mu_1 = \mu_2$ (σ_1^2, σ_2^2 已知)	$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	$N(0, 1)$
	$\mu_1 = \mu_2$ (σ_1^2, σ_2^2 未知)	$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}}$	$t(n_1 + n_2 - 2)$
	$\sigma_1^2 = \sigma_2^2$ (μ_1, μ_2 未知)	$F = \frac{S_1^2}{S_2^2}$	$F(n_1 - 1, n_2 - 1)$

1.5 方差分析与试验设计初步

方差分析(ANOVA)是研究一个或多个自变量(因素)对因变量(响应)的影响,并比较不同因素水平下因变量均值是否存在显著差异的统计方法。它广泛应用于实验设计、质量控制、农业试验、医学研究等领域。方差分析的基本思想是将总变异分解为组间变异和组内变异,通过比较组间变异与组内变异的比值(即F统计量)来判断因素对因变量的影响是否显著。本章将介绍方差分析的基本原理、步骤以及试验设计初步知识,并探讨其与数值型因变量之间的关系。

1.5.1 方差分析概述

首先从一个例子说起:

例 1.17 某环境产品销售量数据表(见表 1.17)。该表记录了在不同广告类型(A₁, A₂, A₃)下,产品在四个季度(一、二、三、四)的销售量。假设销售量受广告类型和季度的影响,试分析广告类型对销售量的影响是否显著。

表 1.17 某环境产品销售量数据表

广告类型	季度				\bar{x}
	一	二	三	四	
A ₁	163	176	170	185	173
A ₂	184	198	179	190	188
A ₃	206	191	218	224	210

从表中可以看出,广告类型 A₁ 强调环保, A₂ 强调绿色, A₃ 强调低碳。假设销售量受广告类型和季度的影响,试分析广告类型对销售量的影响是否显著。若有影响,哪种广告内容比较好?

首先,我们计算每个广告类型在四个季度的销售量均值,并计算总均值。从表中可以看出,广告类型 A₁ 的销售量均值为 173, A₂ 的销售量均值为 188, A₃ 的销售量均值为 210。总均值为 187。接下来,我们将进行方差分析,以判断广告类型对销售量的影响是否显著。

环境统计分析

因此, 在方差分析中, 对于 k 个因素, 每一因素 i 具有 n_i 个水平, 记为 A_1, A_2, \dots, A_{n_i} , 其中 A_1, A_2, \dots, A_{n_i} 表示, 从 k 个因素中, 任意取出一个因素, 可以相等也可以不等。

1.5.2.2 分析步骤

方差分析步骤(贾俊平, 2005):

1. 提出假设

在方差分析中, 提出原假设和备择假设, 原假设和备择假设

$H_0: \mu_1 = \mu_2 = \dots = \mu_k; H_1: \mu_1, \mu_2, \dots, \mu_k$ 不全相等

其中 $\mu_1, \mu_2, \dots, \mu_k$ 表示 k 个因素中, 任意取出一个因素, 可以相等也可以不等, 并不意味着所有的均值都不相等。

2. 构造检验的统计量

(1) 计算因素各水平的均值。

在方差分析中, 计算因素各水平的均值, 记为 \bar{x}_i , 其中 \bar{x}_i 表示, 从 k 个因素中, 任意取出一个因素, 可以相等也可以不等, 并不意味着所有的均值都不相等。

$$\bar{x}_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i} \quad (i=1, 2, \dots, k)$$

其中 x_{ij} 表示, 从 k 个因素中, 任意取出一个因素, 可以相等也可以不等, 并不意味着所有的均值都不相等。

(2) 计算全部观测值的总平均值。

$$\bar{\bar{x}} = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} x_{ij}}{n} = \frac{\sum_{i=1}^k n_i \bar{x}_i}{n} \quad \left(n = \sum_{i=1}^k n_i \right)$$

其中 $\bar{\bar{x}}$ 表示, 从 k 个因素中, 任意取出一个因素, 可以相等也可以不等, 并不意味着所有的均值都不相等。

$$SS_T = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{\bar{x}})^2 \quad (1.28)$$

其中 SS_T 表示, 从 k 个因素中, 任意取出一个因素, 可以相等也可以不等, 并不意味着所有的均值都不相等。

表 1.19 单因素方差分析表

方差来源	平方和	自由度	均方	F 值	F 临界值
组间	SS_A	$k-1$	$MS_A = \frac{SS_A}{k-1}$	$F = \frac{MS_A}{MS_E}$	$F_{\alpha}(k-1, n-k)$
组内	SS_E	$n-k$	$MS_E = \frac{SS_E}{n-k}$		
总和	SS_T	$n-1$			

1.5.2.3 案例

例 1.18 某厂生产 5 种型号的隔音材料, 其噪声衰减率如表 1.20 所示, 问不同型号的隔音材料, 其噪声衰减率是否有明显影响?

表 1.20 不同类型隔音材料的噪声去除效果

材料	噪声衰减率				样本量
A_1	0.110	0.142	0.144		3
A_2	0.152	0.150	0.156	0.154	4
A_3	0.160	0.158	0.163	0.161	4
A_4	0.175	0.173			2
A_5	0.180	0.181	0.182	0.186	4

解 由题意知, 本例为单因素方差分析, 且为均衡设计, 故可用单因素方差分析。将有关统计量列入表 1.21 中。

表 1.21 不同类型隔音材料的样本统计量

材料	A_1	A_2	A_3	A_4	A_5	合计
n	3	4	4	2	4	17
\bar{x}	0.142	0.153	0.161	0.174	0.183	0.162

$$SS_T = 0.003\ 616$$

$$SS_A = 0.003\ 583$$

$$SS_F = SS_T - SS_A = 0.000\ 063$$

方差分析过程见表 1.22。

表 1.22 不同类型隔音材料噪声衰减效果的方差分析

方差来源	平方和	自由度	均方	F 值	$F_{0.05}$	$F_{0.01}$
组间	0.003 583	4	0.000 896	170.61	3.26	5.41
组内	0.000 063	12	0.000 005 25			
总计	0.003 646	16				

由表 1.22 可知, 不同材料噪声衰减效果有显著差异, 且隔音材料对噪声衰减率都有明显影响。

1.5.3 双因素方差分析

方差分析按因素个数可分为单因素方差分析和双因素方差分析。双因素方差分析是指同时考虑两个因素对试验结果的影响的方差分析。双因素方差分析按因素间是否存在交互作用, 可分为无交互作用的双因素方差分析和有交互作用的双因素方差分析。有交互作用的双因素方差分析, 是指两个因素对试验结果的影响不是独立的, 而是相互影响的。有交互作用的双因素方差分析, 是指两个因素对试验结果的影响不是独立的, 而是相互影响的。有交互作用的双因素方差分析, 是指两个因素对试验结果的影响不是独立的, 而是相互影响的。

1.5.3.1 无交互作用的双因素方差分析

无交互作用的双因素方差分析, 是指两个因素对试验结果的影响是独立的, 互不影响的。无交互作用的双因素方差分析, 是指两个因素对试验结果的影响是独立的, 互不影响的。无交互作用的双因素方差分析, 是指两个因素对试验结果的影响是独立的, 互不影响的。无交互作用的双因素方差分析, 是指两个因素对试验结果的影响是独立的, 互不影响的。

值 $x_{11}, \dots, x_{1r}, x_{21}, \dots, x_{2r}, \dots, x_{k1}, \dots, x_{kr}$ 是独立随机变量, 且 x_{ij} 服从正态分布, 方差为 σ^2 , 且 x_{ij} 与 $x_{i'j'}$ 独立 ($i \neq i'$ 或 $j \neq j'$)。总体均值 $\mu_1, \mu_2, \dots, \mu_k$ 未知, 且 $\mu_1, \mu_2, \dots, \mu_k$ 不全相等。从总体中抽取容量为 r 的独立随机样本 $x_{11}, \dots, x_{1r}, x_{21}, \dots, x_{2r}, \dots, x_{k1}, \dots, x_{kr}$ 。

表 1.23 双因素方差分析的数据结构

行因素(i)	列因素(j)				平均值 $\bar{x}_{.j}$
	列 1	列 2	...	列 r	
行 1	x_{11}	x_{12}	...	x_{1r}	$\bar{x}_{1.}$
行 2	x_{21}	x_{22}	...	x_{2r}	$\bar{x}_{2.}$
...
行 k	x_{k1}	x_{k2}	...	x_{kr}	$\bar{x}_{k.}$
平均值 $\bar{x}_{i.}$	$\bar{x}_{.1}$	$\bar{x}_{.2}$...	$\bar{x}_{.r}$	\bar{x}

1. 计算行、列因素的平均值 $\bar{x}_{i.}$ 和 $\bar{x}_{.j}$ 以及总平均值 \bar{x} 。其计算公式为:

$$\bar{x}_{i.} = \frac{\sum_{j=1}^r x_{ij}}{r} \quad (i=1, 2, \dots, k)$$

同理可求得列因素的平均值 $\bar{x}_{.j}$ 。其计算公式为:

$$\bar{x}_{.j} = \frac{\sum_{i=1}^k x_{ij}}{k} \quad (j=1, 2, \dots, r)$$

总平均值 \bar{x} 的计算公式为:

$$\bar{x} = \frac{\sum_{i=1}^k \sum_{j=1}^r x_{ij}}{kr}$$

2. 分析步骤

双因素方差分析, 包括行、列两因素总体的均值假设、方差检验、统计量、决策分析等步骤。

(1) 提出假设

与单因素方差分析类似, 双因素方差分析对两因素分别提出原假设。

对行因素提出的假设为:

$$H_0: \mu_1 = \mu_2 = \dots = \mu_i = \dots = \mu_k; H_1: \mu_i (i=1, 2, \dots, k) \text{ 不全相等}$$

对列因素提出的假设为:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_j = \cdots = \mu_r; H_1: \mu_j (j=1, 2, \cdots, r) \text{ 不全相等}$$

(2) 构造检验统计量

总平方和 SS_T 是全部数据 x_{ij} 与总平均值 \bar{x} 的离差平方和, 记为 SS_T , 即:

$$\begin{aligned} SS_T &= \sum_{i=1}^k \sum_{j=1}^r (x_{ij} - \bar{x})^2 \\ &= \sum_{i=1}^k \sum_{j=1}^r (x_{ij} - \bar{x}_i + \bar{x}_i - \bar{x})^2 + \sum_{j=1}^r \sum_{i=1}^k (x_{ij} - \bar{x}_i + \bar{x}_i - \bar{x})^2 + \sum_{i=1}^k \sum_{j=1}^r (x_{ij} - \bar{x}_i + \bar{x}_i - \bar{x})^2 \\ &= r \sum_{i=1}^k (x_i - \bar{x})^2 + k \sum_{j=1}^r (x_j - \bar{x})^2 + \sum_{i=1}^k \sum_{j=1}^r (x_{ij} - \bar{x}_i + \bar{x}_i - \bar{x})^2 \end{aligned} \quad (1.35)$$

上式右端第一项 $r \sum_{i=1}^k (x_i - \bar{x})^2$ 为行因素平方和, 记为 SS_R , 即:

$$SS_R = r \sum_{i=1}^k (\bar{x}_i - \bar{x})^2 \quad (1.36)$$

第二项 $k \sum_{j=1}^r (x_j - \bar{x})^2$ 为列因素平方和, 记为 SS_C , 即:

$$SS_C = k \sum_{j=1}^r (\bar{x}_j - \bar{x})^2 \quad (1.37)$$

第三项 $\sum_{i=1}^k \sum_{j=1}^r (x_{ij} - \bar{x}_i + \bar{x}_i - \bar{x})^2$ 为交互作用平方和, 记为 SS_E , 即:

$$SS_E = \sum_{i=1}^k \sum_{j=1}^r (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{x})^2 \quad (1.38)$$

上述各平方和的关系为:

$$SS_T = SS_R + SS_C + SS_E \quad (1.39)$$

由上式可知, 总平方和 SS_T 等于行因素平方和 SS_R 、列因素平方和 SS_C 与交互作用平方和 SS_E 之和。因此, 只要计算出 SS_T 、 SS_R 和 SS_C , 即可求出 SS_E 。交互作用平方和 SS_E 的自由度为 $(k-1)(r-1)$ 。

为了构造检验统计量, 需要计算下列各均方:

行因素的均方, 记为 MS_R :

$$MS_R = \frac{SS_R}{k-1} \quad (1.40)$$

列因素的均方, 记为 MS_c :

$$MS_c = \frac{SS_c}{r-1} \quad (1.41)$$

随机误差的均方, 记为 MS_e :

$$MS_e = \frac{SS_e}{(k-1)(r-1)} \quad (1.42)$$

$$F_R = \frac{MS_R}{MS_e} \sim F(k-1, (k-1)(r-1)) \quad (1.43)$$

为了检验列因素的影响是否显著, 采用下面的统计量:

$$F_c = \frac{MS_c}{MS_e} \sim F(r-1, (k-1)(r-1)) \quad (1.44)$$

(3) 统计决策

查 F 分布表, 得 $F_{\alpha}(k-1, (k-1)(r-1))$ 和 $F_{\alpha}(r-1, (k-1)(r-1))$ 相应的临界值 F_{α} , 然后将 F_R 、 F_c 与 F_{α} 进行比较:

若 $F_R > F_{\alpha}$, 则拒绝 H_0 , 认为行因素 A 对试验结果有显著影响, 表明它们之间的差异是显著的。

若 $F_c > F_{\alpha}$, 则拒绝 H_0 , 认为列因素 B 对试验结果有显著影响, 表明它们之间的差异是显著的。

表 1.24 双因素方差分析表的结构

方差来源	平方和	自由度	均方	F 值	F 的临界值
行因素 A	SS_R	$k-1$	$MS_R = \frac{SS_R}{k-1}$	$F_R = \frac{MS_R}{MS_e}$	$F_{\alpha}(k-1, (k-1)(r-1))$
列因素 B	SS_c	$r-1$	$MS_c = \frac{SS_c}{r-1}$	$F_c = \frac{MS_c}{MS_e}$	$F_{\alpha}(r-1, (k-1)(r-1))$
误差	SS_e	$(k-1)(r-1)$	$MS_e = \frac{SS_e}{(k-1)(r-1)}$		
总和	SS_T	$kr-1$			

3. 案例

例 1.19 某工厂生产一种产品, 为了降低成本, 考虑对材料、工艺、设备三个因素进行试验, 每个因素有三个水平, 记为 A、B、C, 甲、乙、丙, 丁、戊、己, 共 27 种组合。试验结果如下表所示, 试分析各因素对成本的影响。

表 1.25

数据表

原料来源地(A)	环保产品合格率		
	B ₁ (现用量)	B ₂ (增加5%)	B ₃ (增加8%)
甲(A ₁)	59	70	66
乙(A ₂)	63	74	70
丙(A ₃)	61	66	71

解 1. $x_i = 0, 1, 2, \dots, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198, 199, 200, 201, 202, 203, 204, 205, 206, 207, 208, 209, 210, 211, 212, 213, 214, 215, 216, 217, 218, 219, 220, 221, 222, 223, 224, 225, 226, 227, 228, 229, 230, 231, 232, 233, 234, 235, 236, 237, 238, 239, 240, 241, 242, 243, 244, 245, 246, 247, 248, 249, 250, 251, 252, 253, 254, 255, 256, 257, 258, 259, 260, 261, 262, 263, 264, 265, 266, 267, 268, 269, 270, 271, 272, 273, 274, 275, 276, 277, 278, 279, 280, 281, 282, 283, 284, 285, 286, 287, 288, 289, 290, 291, 292, 293, 294, 295, 296, 297, 298, 299, 300, 301, 302, 303, 304, 305, 306, 307, 308, 309, 310, 311, 312, 313, 314, 315, 316, 317, 318, 319, 320, 321, 322, 323, 324, 325, 326, 327, 328, 329, 330, 331, 332, 333, 334, 335, 336, 337, 338, 339, 340, 341, 342, 343, 344, 345, 346, 347, 348, 349, 350, 351, 352, 353, 354, 355, 356, 357, 358, 359, 360, 361, 362, 363, 364, 365, 366, 367, 368, 369, 370, 371, 372, 373, 374, 375, 376, 377, 378, 379, 380, 381, 382, 383, 384, 385, 386, 387, 388, 389, 390, 391, 392, 393, 394, 395, 396, 397, 398, 399, 400, 401, 402, 403, 404, 405, 406, 407, 408, 409, 410, 411, 412, 413, 414, 415, 416, 417, 418, 419, 420, 421, 422, 423, 424, 425, 426, 427, 428, 429, 430, 431, 432, 433, 434, 435, 436, 437, 438, 439, 440, 441, 442, 443, 444, 445, 446, 447, 448, 449, 450, 451, 452, 453, 454, 455, 456, 457, 458, 459, 460, 461, 462, 463, 464, 465, 466, 467, 468, 469, 470, 471, 472, 473, 474, 475, 476, 477, 478, 479, 480, 481, 482, 483, 484, 485, 486, 487, 488, 489, 490, 491, 492, 493, 494, 495, 496, 497, 498, 499, 500, 501, 502, 503, 504, 505, 506, 507, 508, 509, 510, 511, 512, 513, 514, 515, 516, 517, 518, 519, 520, 521, 522, 523, 524, 525, 526, 527, 528, 529, 530, 531, 532, 533, 534, 535, 536, 537, 538, 539, 540, 541, 542, 543, 544, 545, 546, 547, 548, 549, 550, 551, 552, 553, 554, 555, 556, 557, 558, 559, 560, 561, 562, 563, 564, 565, 566, 567, 568, 569, 570, 571, 572, 573, 574, 575, 576, 577, 578, 579, 580, 581, 582, 583, 584, 585, 586, 587, 588, 589, 590, 591, 592, 593, 594, 595, 596, 597, 598, 599, 600, 601, 602, 603, 604, 605, 606, 607, 608, 609, 610, 611, 612, 613, 614, 615, 616, 617, 618, 619, 620, 621, 622, 623, 624, 625, 626, 627, 628, 629, 630, 631, 632, 633, 634, 635, 636, 637, 638, 639, 640, 641, 642, 643, 644, 645, 646, 647, 648, 649, 650, 651, 652, 653, 654, 655, 656, 657, 658, 659, 660, 661, 662, 663, 664, 665, 666, 667, 668, 669, 670, 671, 672, 673, 674, 675, 676, 677, 678, 679, 680, 681, 682, 683, 684, 685, 686, 687, 688, 689, 690, 691, 692, 693, 694, 695, 696, 697, 698, 699, 700, 701, 702, 703, 704, 705, 706, 707, 708, 709, 710, 711, 712, 713, 714, 715, 716, 717, 718, 719, 720, 721, 722, 723, 724, 725, 726, 727, 728, 729, 730, 731, 732, 733, 734, 735, 736, 737, 738, 739, 740, 741, 742, 743, 744, 745, 746, 747, 748, 749, 750, 751, 752, 753, 754, 755, 756, 757, 758, 759, 760, 761, 762, 763, 764, 765, 766, 767, 768, 769, 770, 771, 772, 773, 774, 775, 776, 777, 778, 779, 780, 781, 782, 783, 784, 785, 786, 787, 788, 789, 790, 791, 792, 793, 794, 795, 796, 797, 798, 799, 800, 801, 802, 803, 804, 805, 806, 807, 808, 809, 810, 811, 812, 813, 814, 815, 816, 817, 818, 819, 820, 821, 822, 823, 824, 825, 826, 827, 828, 829, 830, 831, 832, 833, 834, 835, 836, 837, 838, 839, 840, 841, 84$

表 1.26

双因素方差分析表

方差来源	平方和	自由度	均方	F 值	$F_{0.05}(2, 4)$
因素 A	26	2	13	1.86	6.94
因素 B	146	2	73	10.43	6.94
误差	28	4	7		
总和	200	8			

使用量对合格率有显著影响。

1.5.3.2 有交互作用的双因素方差分析

这就是有交互作用的双因素方差分析。

$$f(x) = \begin{cases} x^2 & x \in (-\infty, -1] \\ x^2 + 2x & x \in (-1, 0] \\ x^2 - 2x & x \in (0, 1] \\ x^2 & x \in (1, +\infty) \end{cases}$$

表 1.27 有交互作用的双因素试验数据表

行因素(i)	列因素(j)				平均值 \bar{x}_i
	列 1	列 2	...	列 r	
行 1	$x_{11}, \dots, x_{1m_{11}}, x_{121}, \dots, x_{12m_{12}}, \dots, x_{1r}, \dots, x_{1rm_{1r}}$				\bar{x}_1
行 2	$x_{21}, \dots, x_{2m_{21}}, x_{221}, \dots, x_{22m_{22}}, \dots, x_{2r}, \dots, x_{2rm_{2r}}$				\bar{x}_2
...					
行 k	$x_{k1}, \dots, x_{km_{k1}}, x_{k21}, \dots, x_{k2m_{k2}}, \dots, x_{kr}, \dots, x_{krm_{kr}}$				\bar{x}_k
平均值 \bar{x}_j	\bar{x}_1	\bar{x}_2	...	\bar{x}_r	$\bar{\bar{x}}$

1. 行因素和列因素: 行因素是指试验中, 除列因素外, 其他因素都保持不变的条件下, 改变行因素的水平, 所得到的试验结果。列因素是指试验中, 除行因素外, 其他因素都保持不变的条件下, 改变列因素的水平, 所得到的试验结果。

值, $n = \sum_{i=1}^k \sum_{j=1}^r m_{ij}$ 。

各平方和的计算公式如下:

总偏差平方和(SS_T):

$$SS_T = \sum_{i=1}^k \sum_{j=1}^r \sum_{t=1}^{m_{ij}} (x_{ijt} - \bar{\bar{x}})^2 \quad (1.45)$$

行变量平方和(SS_R):

$$SS_R = \sum_{i=1}^k m_i (x_i - \bar{\bar{x}})^2 \quad (1.46)$$

列变量平方和(SS_C):

$$SS_C = \sum_{j=1}^r m_j (x_j - \bar{\bar{x}})^2 \quad (1.47)$$

交互作用平方和(SS_{RC}):

$$SS_{RC} = \sum_{i=1}^k \sum_{j=1}^r m_{ij} (x_{ij} - x_i - \bar{x}_j + \bar{\bar{x}})^2 \quad (1.48)$$

误差平方和(SS_E):

$$SS_E = SS_T - SS_R - SS_C - SS_{RC} \quad (1.49)$$

表 1.23 有交互作用的双因素方差分析表的结构

方差来源	平方和	自由度	均方	F 值	F 的临界值
行因素	SS_R	$k-1$	$MS_R = \frac{SS_R}{k-1}$	$F_R = \frac{MS_R}{MS_E}$	$F_\alpha(k-1, n-kr)$
列因素	SS_C	$r-1$	$MS_C = \frac{SS_C}{r-1}$	$F_C = \frac{MS_C}{MS_E}$	$F_\alpha(r-1, n-kr)$
交互作用	SS_{RC}	$(k-1)(r-1)$	$MS_{RC} = \frac{SS_{RC}}{(k-1)(r-1)}$	$F_{RC} = \frac{MS_{RC}}{MS_E}$	$F_\alpha((k-1)(r-1), n-kr)$
误差	SS_E	$n-kr$	MS_E		
总和	SS_T	$n-1$			

1.5.4 试验设计初步

试验设计是统计学的一个重要分支，是数据分析方法、统计推断方法的基础，本书主要介绍试验设计的一些基本知识。

1.5.4.1 完全随机化设计

完全随机化设计是指将试验对象随机地分成若干组，每组接受不同的处理，然后比较各组的结果。这种设计适用于处理组和对照组的比较。例如，在比较两种药物的疗效时，可以将患者随机分为两组，一组接受药物治疗，另一组接受安慰剂，然后比较两组的疗效。完全随机化设计的特点是操作简单、易于实施，但可能存在个体差异的影响。

例 1.20 某工厂生产的产品，其质量受温度、湿度、时间等因素影响。为了研究这些因素对产品质量的影响，设计了一个完全随机化试验。将产品随机分为两组，一组在高温下生产，另一组在低温下生产。然后比较两组产品的质量。试验结果表明，高温生产的产品质量较差，而低温生产的产品质量较好。这就是试验设计的过程。

完全随机化设计的特点是操作简单、易于实施，但可能存在个体差异的影响。为了提高试验的精度，可以采用随机化设计。随机化设计是指将试验对象随机地分配到不同的处理组，然后比较各组的结果。这种设计可以有效地消除个体差异的影响，提高试验的精度。例如，在比较两种药物的疗效时，可以将患者随机分为两组，一组接受药物治疗，另一组接受安慰剂，然后比较两组的疗效。随机化设计的特点是操作简单、易于实施，但可能存在个体差异的影响。

环境统计分析

重复是指在一个试验中，同一处理重复进行“多次”，构成了“重复组”，由于重复组的存在，使得试验结果更加准确。重复组在田间试验中通常有 3 种形式：完全重复、局部重复和不完全重复。完全重复是指每个处理组在试验中重复进行多次，且每次重复都是在同一块地上进行的。局部重复是指每个处理组在试验中重复进行多次，但每次重复是在不同的地块上进行的。不完全重复是指每个处理组在试验中重复进行多次，但每次重复是在不同的地块上进行的，且每次重复的重复次数不同。

表 1.29 3 种肥料在 12 个地块上的产量数据

类 型	产 量			
无机肥	368	349	351	342
普通有机肥	336	383	370	357
氨基酸复合肥	351	348	336	331

对表 1.29 的数据，用方差分析方法，求出 3 种肥料在 12 个地块上的产量数据，用方差分析方法进行分析。表 1.30 给出了对数据的分析结果。

表 1.30 3 种肥料的方差分析表

方差来源	平方和	自由度	均方	F 值	F _α
组间	2 186	2	1 093	8.42	4.26
组内	1 168	9	130		
总和	3 354	11			

由表 1.30 可知，3 种肥料在 12 个地块上的产量数据，表 1.30 肥料和产量数据如下：

1.5.4.2 因子设计

因子设计是指在一个试验中，将不同的处理因素（因子）进行组合，以研究其对试验结果的影响。因子设计通常分为完全因子设计和部分因子设计。完全因子设计是指每个处理因素的所有水平都进行组合。部分因子设计是指只进行部分组合。因子设计通常用于研究多个因素对试验结果的影响，以及因素之间的交互作用。

例 1.21 假设在一个试验中，研究 3 种小麦品种（甲、乙、丙）在不同肥料（无机肥、普通有机肥、氨基酸复合肥）下的产量数据，见表 1.31。

假设我们研究 3 种小麦品种（甲、乙、丙）在不同肥料（无机肥、普通有机肥、氨基酸复合肥）下的产量数据，见表 1.31。假设我们研究 3 种小麦品种（甲、乙、丙）在不同肥料（无机肥、普通有机肥、氨基酸复合肥）下的产量数据，见表 1.31。

表 1.31 肥料类型和小麦品种的因子试验数据

肥料类型	小麦品种	
	甲	乙
人机肥	81	89
	82	92
	79	87
	81	85
	78	86
普通有机肥	71	77
	72	81
	72	77
	66	73
	72	79
氨基酸复合肥	76	89
	79	87
	77	84
	76	87
	78	87

表 1.32 小麦品种和肥料类型因子试验的方差分析表

方差来源	平方和	自由度	均方	F 值	F _α
行因素	560	2	280	54.37	3.40
列因素	480	1	480	93.20	4.26
交互作用	10.4	2	5.2	1.01	3.40
误差	123.6	24	5.15		
总和	1 174	29			

从表 1.32 中可以看出，行因素和列因素的 F 值均大于 F_α，说明行因素和列因素对小麦产量的影响是显著的。交互作用的 F 值小于 F_α，说明交互作用对小麦产量的影响是不显著的。因此，在分析小麦产量时，应主要考虑行因素和列因素的影响，而无需考虑交互作用的影响。

1. 举例说明常用的几个统计量。
 2. 详述假设检验的步骤。
 3. 详述频率直方图和累积频率图的步骤。

【思考题 1】

1. 举例说明常用的几个统计量。

大小和意义。

3. 详述假设检验的步骤

4. 详述频率直方图和累积频率图的步骤。

净重(单位: g)数据如下:

497, 506, 518, 524, 488, 517, 510, 515, 516

若取显著性水平 $\alpha=0.01$, 问这台包装机工作是否正常?

净重是否有显著影响?

气中飘尘含量的时空差异是否显著。

表 1.33

某市大气飘尘监测结果 单位: mg/m^3

	春季	夏季	秋季	冬季
市中心区	0.620	0.420	0.880	1.20
近郊区	0.614	0.475	0.667	1.15
远郊区	0.379	0.200	0.540	0.94

【参考文献】

- [7] 贾俊平. 统计学[M]. 北京: 清华大学出版社, 2005.

第2章 环境一元线性回归分析

环境一元线性回归分析是环境统计学中应用最广泛的一种统计方法。它主要用于研究两个变量之间的线性关系。在环境科学中，许多环境指标之间存在线性关系，例如，大气污染物的浓度与气象条件、水体的污染程度与工业排放量等。通过建立一元线性回归模型，可以定量地描述这种关系，并用于预测和决策。本章将详细介绍一元线性回归的基本原理、模型建立、参数估计、显著性检验、误差估计以及在实际环境问题中的应用。

解决实际环境问题

本章的主要内容是：

- 一元线性回归的建模原理；
- 模型参数的最小二乘估计；
- 线性回归方程的显著性检验；
- 线性回归式的误差估计；
- 可化为一元线性回归的曲线回归；
- 环境应用。

2.1 一元线性回归模型

2.1.1 变量间的统计关系

在环境科学中，许多环境指标之间存在统计关系。这种关系可以是线性的，也可以是曲线的。一元线性回归模型是研究两个变量之间线性关系的一种统计方法。它假设一个变量（因变量）与另一个变量（自变量）之间存在线性关系，并通过最小二乘法来估计回归方程的参数。在实际应用中，一元线性回归模型广泛应用于环境质量的评估、污染源的追踪以及环境政策的制定等方面。例如，通过建立大气污染物浓度与气象条件之间的线性回归模型，可以预测不同气象条件下的污染物浓度，从而为环境管理部门提供决策依据。

$$f_1, \dots, f_r \in R, \quad f_1^2 + \dots + f_r^2 \neq 0. \quad \text{K} \text{ (} f_1^2 + \dots + f_r^2 \text{)} \text{ is a prime ideal, } f_1^2 + \dots + f_r^2 \in \text{K}.$$

样, 一元线性回归模型分析一元线性回归模型分析之。

2.1.2 一元线性回归模型

“回归”一词, 最初由英国统计学家高尔顿(G. Galton)在研究人体身高时提出。高尔顿在研究父亲的身高与儿子的身高之间的关系时, 发现了一个有趣的现象: 如果父亲的身高高于平均身高, 那么儿子的身高也会高于平均身高, 但不会高得离谱; 反之, 如果父亲的身高低于平均身高, 那么儿子的身高也会低于平均身高, 但不会低得离谱。高尔顿将这种现象称为“回归”, 即儿子的身高会“回归”到平均身高。高尔顿的这一发现为后来的统计学家提供了重要的启示, 即变量之间存在一种内在的、稳定的、可预测的关系。这种关系可以用数学模型来描述, 即回归模型。回归模型是统计建模中最基本、最重要的模型之一, 广泛应用于各个领域, 如经济学、社会学、医学、生物学等。在环境统计中, 回归模型常用于分析环境因素与污染指标之间的关系, 如温度与溶解氧浓度的关系、流量与溶解氧浓度的关系等。

回归分析是一种统计方法, 用于研究变量之间的关系。回归分析的基本思想是: 假设因变量与自变量之间存在一种线性关系, 通过收集数据, 拟合出一条回归直线, 利用这条直线来估计、预测因变量。回归分析分为一元线性回归和多元线性回归。一元线性回归是指只有一个自变量的情况, 多元线性回归是指有多个自变量的情况。回归分析的结果通常用回归系数、截距、决定系数等指标来描述。回归分析在环境统计中具有重要的应用价值, 可以帮助研究人员了解环境因素对污染指标的影响, 为环境管理和污染治理提供科学依据。

在回归分析中, 回归系数是一个重要的参数, 它反映了自变量对因变量的影响程度。回归系数的估计通常采用最小二乘法。最小二乘法的基本思想是: 寻找一条直线, 使得所有数据点到这条直线的距离平方和最小。最小二乘法是一种经典的统计方法, 具有简单易用、计算方便等优点。在环境统计中, 最小二乘法常用于估计回归系数, 分析环境因素与污染指标之间的关系。例如, 可以通过最小二乘法估计溶解氧浓度与流量之间的关系, 了解流量对溶解氧浓度的影响。

例 2.1 某河流在不同流量下的溶解氧浓度数据如下表所示, 现拟建立 8 组数据, 如表 2.1 所示。

表 2.1 河流中溶解氧浓度

流动时间 x/d	0.5	1.0	1.6	1.8	2.6	3.2	3.8	4.7
溶解氧浓度 y	0.28	0.29	0.29	0.18	0.17	0.18	0.10	0.12

由表 2.1 可知, 该组数据为一元线性回归模型, 拟建立一元线性回归模型, 分析

在《几何原本》中， $\sqrt{2}$ 和 $\sqrt{5}$ 是无理数，所以，把无理数“各”看做一条直线上的无数条射线，那么，直线上全有无数条射线。

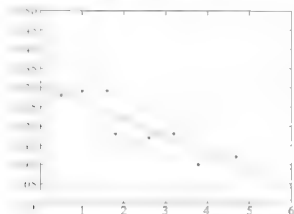


图 2-1 溶解氧浓度随时间变化曲线

[illegible]

元线性回归模型为:

$$y = a + bx + \epsilon$$

[illegible]

假定 $E(\epsilon) = 0$, 有回归函数:

$$\mu(x) = E(y) = a + bx \quad (2.1)$$

[illegible]

2.1.3 最小二乘法估计

的实际值应为:

进行 n 次独立观测, 假设实测数据为 $(x_i, y_i) (i=1, 2, \dots, n)$, 即:

$$\begin{aligned} x_1, x_2, x_3, \dots, x_n \\ y_1, y_2, y_3, \dots, y_n \end{aligned} \quad (2.2)$$

下最小,

$$\epsilon_i = y_i - \mu(x_i) \quad (i=1, 2, \dots, n)$$

上式也可以写为:

$$y_i = a + bx_i + \epsilon_i \quad (i=1, 2, \dots, n) \quad (2.3)$$

$(i=1, 2, \dots, n)$ 且期望 $E(\epsilon_i) = 0$, $D(\epsilon_i) = \sigma^2$, $\text{Cov}(\epsilon_i, x_i) = 0$ ($i=1, 2, \dots, n$), 所

$$Q = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n [y_i - (a + bx_i)]^2 \quad (2.4)$$

$$\begin{aligned} \frac{\partial Q}{\partial a} &= -2 \sum_{i=1}^n [y_i - (a + bx_i)] = 0 \\ \frac{\partial Q}{\partial b} &= -2 \sum_{i=1}^n [y_i - (a + bx_i)] x_i = 0 \end{aligned} \quad (2.5)$$

解正规方程组可得:

$$\hat{b} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{a} = \bar{y} - \hat{b}\bar{x} \quad (2.6)$$

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}), \quad \sum_{i=1}^n (x_i - \bar{x})^2, \quad \sum_{i=1}^n (y_i - \bar{y})^2$$

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}), \quad L_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}), \quad L_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2, \quad L_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$$

则

$$\hat{b} = \frac{L_{xy}}{L_{xx}}, \quad \hat{a} = \bar{y} - \hat{b}\bar{x} \quad (2.7)$$

取 $\hat{a} + \hat{b}x$ 作为 $\mu(x) = a + bx$ 的估计, 记 $\hat{y} = \hat{a} + \hat{b}x$.

squares estimators, (OLSE).

根据公式, 求例 2.1, 可得:

$$\bar{x} = \frac{1}{8} \sum_{i=1}^8 x_i = 2.4, \quad \bar{y} = \frac{1}{8} \sum_{i=1}^8 y_i = 1.61 \times \frac{1}{8} = 0.20$$

$$L_{xy} = 14.5000, \quad L_{xx} = 0.6840, \quad L_{yy} = 0.0407$$

由此可得:

$$\hat{b} = \frac{L_{xy}}{L_{xx}} = -0.0472, \quad \hat{a} = \bar{y} - \hat{b}\bar{x} = 0.3145$$

$$\hat{y} = \hat{a} + \hat{b}x = 0.3145 - 0.0472x$$

2.2 线性回归方程的显著性检验

进行统计推断

$$\text{则: } F = \frac{S_{\text{回}}}{S_{\text{残}}(n-2)} = \frac{0.0323}{0.00846} = 23.0714$$

当 $\alpha=0.05$ 时, 查 F 分布表, 得到 $F_{\alpha}(1, n-2)=5.9900$, 由于 $F=23.0714 > 5.9900 = F_{\alpha}(1, n-2)$, 所以 $\hat{y} = \hat{a} + \hat{b}x = 0.3145 + 0.0472x$ 线性关系显著。

2.2.2 相关系数检验法

检验的方法 相关系数法

$$\begin{aligned} \text{因为 } S_{\text{回}} &= \sum (y_i - \hat{y}_i)^2 = \sum [y_i - (\hat{a} + \hat{b}x_i)]^2 \\ &= \sum [y_i - (y - \hat{b}x + \hat{b}x_i)]^2 \\ &= \sum [(y_i - y) - \hat{b}(x_i - \bar{x})]^2 \\ &= \sum (y_i - y)^2 - 2\hat{b} \sum (x_i - \bar{x})(y_i - \bar{y}) + \hat{b}^2 \sum (x_i - \bar{x})^2 \\ &= \sum (y_i - y)^2 - 2\hat{b} \sum (x_i - \bar{x})\bar{y} + \hat{b}^2 \sum (x_i - \bar{x})^2 \\ &= \sum (y_i - \bar{y})^2 - \hat{b}^2 \sum (x_i - \bar{x})^2 = L_{yy} - \hat{b}^2 L_{xx} \end{aligned}$$

$$r^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \hat{b}^2 \sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{S_{\text{残}}}{L_{yy}}$$

在 x_0 处, 估计值 \hat{y}_0 与 y_0 之差, 称为估计值与真值之差, 记为 $\hat{y}_0 - y_0$ 。从统计推断的角度看, 估计值 \hat{y}_0 与真值 y_0 之差, 称为估计误差。因此, 估计系数 b 与截距 a 的估计误差, 称为估计误差。估计误差平方和为 $\sum_{i=1}^n (\hat{y}_i - y_i)^2$, 记为 L_D 来解释。

在多元线性回归中, 估计值 \hat{y}_0 与真值 y_0 之差, 称为估计误差。估计误差平方和, 记为 L_D 。在多元线性回归中, 估计值 \hat{y}_0 与真值 y_0 之差, 称为估计误差。估计误差平方和, 记为 L_D 。在多元线性回归中, 估计值 \hat{y}_0 与真值 y_0 之差, 称为估计误差。估计误差平方和, 记为 L_D 。

2.3 线性回归式的误差估计

2.3.1 线性回归式的误差估计

在多元线性回归中, 估计值 \hat{y}_0 与真值 y_0 之差, 称为估计误差。估计误差平方和, 记为 L_D 。在多元线性回归中, 估计值 \hat{y}_0 与真值 y_0 之差, 称为估计误差。估计误差平方和, 记为 L_D 。在多元线性回归中, 估计值 \hat{y}_0 与真值 y_0 之差, 称为估计误差。估计误差平方和, 记为 L_D 。

在多元线性回归中, 估计值 \hat{y}_0 与真值 y_0 之差, 称为估计误差。估计误差平方和, 记为 L_D 。在多元线性回归中, 估计值 \hat{y}_0 与真值 y_0 之差, 称为估计误差。估计误差平方和, 记为 L_D 。

$$\frac{\hat{y}_0 - y_0}{\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim t(n-2)$$

在多元线性回归中, 估计值 \hat{y}_0 与真值 y_0 之差, 称为估计误差。估计误差平方和, 记为 L_D 。

$$P\left\{ \left| \frac{\hat{y}_0 - y_0}{\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \right| \leq t_{\alpha}(n-2) \right\} = 1 - \alpha$$

在多元线性回归中, 估计值 \hat{y}_0 与真值 y_0 之差, 称为估计误差。估计误差平方和, 记为 L_D 。

$$\hat{\sigma}^2 = \frac{S_{\hat{y}}}{n-2} = \frac{L_D}{n-2}$$

在多元线性回归中, 估计值 \hat{y}_0 与真值 y_0 之差, 称为估计误差。估计误差平方和, 记为 L_D 。

故 $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2$ 。

当 n 较大时, 近似有(卢崇飞等, 1988):

$$\sqrt{n}(\hat{\sigma}^2 - \sigma^2) \xrightarrow{D} N(0, \sigma^4)$$

于是, 近似有:

$$P(\hat{\sigma}^2 - 3\hat{\sigma} < \hat{\sigma}^2 < \hat{\sigma}^2 + 3\hat{\sigma}) \\ 0.99 \approx P(\hat{y}_0 - 3\hat{\sigma} < y_0 < \hat{y}_0 + 3\hat{\sigma})$$

即, 置信方程, 置信度为 0.99 时, 置信区间 $P(\hat{y}_0 - 3\hat{\sigma} < y_0 < \hat{y}_0 + 3\hat{\sigma}) = 0.99$ 的误差可以忽略。

2.3.2 线性回归的步骤

首先, 对给定的具体材料, 现在将一元线性回归分析的具体步骤如下列示:

(1) 设变量 x 和 y 的线性回归方程为: $\hat{y} = \hat{a} + \hat{b}x$ 。

$$b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad a = \bar{y} - b\bar{x}$$

$$b = \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n} (\sum_{i=1}^n x_i) (\sum_{i=1}^n y_i)}{\sum_{i=1}^n x_i^2 - \frac{1}{n} (\sum_{i=1}^n x_i)^2}, \quad a = \bar{y} - b\bar{x} \quad \text{和} \quad t = \frac{\sum_{i=1}^n y_i}{n}$$

$$\sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i^2 - \frac{1}{n} (\sum_{i=1}^n x_i)^2 = \sum_{i=1}^n x_i^2 - \frac{1}{n} (\sum_{i=1}^n x_i)^2, \quad \text{求和时因 } t \text{ 和 } t^2$$

$$\sum_{i=1}^n y_i^2 = \sum_{i=1}^n y_i^2 - \frac{1}{n} (\sum_{i=1}^n y_i)^2$$

(3) 检验回归系数 b 是否为零。

检验 $H_0: b=0$ 成立时, 令:

$$F = \frac{S_{\text{回}}}{S_{\text{残}}/(n-2)} \sim F(1, n-2)$$

若 $F > F_{\alpha}(1, n-2)$, 则在 $1-\alpha$ 水平下拒绝 H_0 ; 若 $F \leq F_{\alpha}(1, n-2)$, 则接受 H_0 。

$$F_{\alpha}(1, n-2) = F_{\alpha}(1, n-2), \quad \text{查表 } H_0: b=0, \text{ 认为一元线性回归方程}$$

1. $F = \frac{S_{\text{回}}}{S_{\text{残}}/(n-2)} > F_{\alpha}(1, n-2)$, 则接受 $H_0: b=0$, 认为一元线性回归方程

又

(4) 求相关系数并作相关性检验。

$$\text{首先求得相关系数 } r = \frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}$$

检验方法如下:

当 $\chi^2 < \chi^2_{\alpha}$ 时, 认为 $\rho = 0$, 即 x 与 y 不存在线性关系;
当 $\chi^2 > \chi^2_{\alpha}$ 时, 认为 $\rho \neq 0$, 即 x 与 y 存在线性关系。

2.4 可化为一元线性回归的曲线回归

在多元回归分析中, 当自变量 x 与因变量 y 之间的关系不是线性关系, 而是某种曲线关系时, 如果能够通过适当的变换, 将这种曲线关系转化为线性关系, 那么就可以利用一元线性回归的方法来求解参数。这种将非线性关系转化为线性关系的方法称为曲线回归。常见的曲线回归方法有: 倒数变换、对数变换、平方根变换、平方变换、取对数变换等。

下面以倒数变换为例, 说明如何将非线性关系转化为线性关系。假设 y 与 x 之间的关系为 $y = \frac{a}{x} + b$, 其中 a 和 b 为待求参数。对等式两边取倒数, 得到 $\frac{1}{y} = \frac{x}{a + bx}$ 。令 $Y = \frac{1}{y}$, $X = \frac{x}{a + bx}$, 则原方程可化为 $Y = X$ 。此时, Y 与 X 之间的关系为线性关系, 可以通过一元线性回归的方法来求解参数 a 和 b 。

下面就列举几个常用的转化方法。

2.4.1 倒数变换

(1) 双曲线(图 2-2)

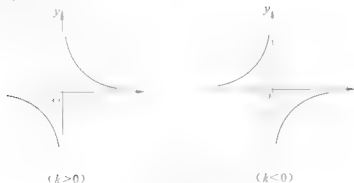


图 2-2 双曲线示意图

反比例函数 $\frac{1}{y} = \frac{k}{x} \quad (k \neq 0, x \neq 0, y \neq 0)$ (2.8)

令 $y' = \frac{1}{y}$, $x' = \frac{1}{x}$; 则得

$$y' = kx'$$

(2) S 型曲线(图 2-3)



图 2-3 S 型曲线示意图

$$y = \frac{a}{a + be^{-x}} \quad (a + be^{-x} \neq 0)$$
 (2.9)

令 $y' = \frac{1}{y}$, $x' = e^{-x}$, 则得

$$y' = a + bx'$$

2.4.2 对数变换

将曲线回归分析转化为线性回归分析。

(1) 指数函数(图 2-4)



图 2-4 指数函数曲线示意图

$$y = ae^{bx} \quad (a > 0) \quad (2.10)$$

对其两边取自然对数, 得 $\ln y = \ln a + bx$

令 $y' = \ln y$, 则得 $y' = \ln a + bx$

(2) 对数函数(图 2-5)



图 2-5 对数函数曲线示意图

$$y = a + b \lg x \quad (x > 0) \quad (2.11)$$

令 $x' = \lg x$, 则得 $y = a + bx'$

(3) 幂函数(图 2-6)

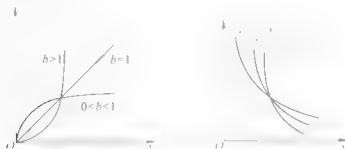


图 2-6 幂函数曲线示意图

$$y = ax^b \quad (a > 0, x > 0) \quad (2.12)$$

对上式两边取对数, 得 $\lg y = \lg a + b \lg x$

令 $y' = \lg y$, $x' = \lg x$, 则得 $y' = \lg a + bx'$

2.4.3 混合变换

在回归分析中, 当遇到非线性化, 且用多种变换未能实现时, 可将函数进行混合变换, 即对函数进行多次变换, 如 $y = a + b \lg x$ 和 $y = ax^b$ 混合, 得到



图 2-7

$$L_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - 20\bar{x} = 75.7575$$

$$L_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - 20\bar{y} = 36.8255$$

$$\hat{a} = \bar{y} - b\bar{x} = 2.6650 - 0.6104 \times 3.3750 = 0.6049$$

所以, 得到的线性回归方程为: $\hat{y} = \hat{a} + \hat{b}x = 0.6049 + 0.6104x$

$$(3) S_E = L_{yy} = 36.8255, S_{\mu} = 28.2263, S_{\mu} = S_E - S_{\mu} = 8.5992$$

$$\text{则 } F = \frac{S_{\mu}}{S_E} \frac{(n-2)}{8.5992/18} = 59.0842$$

$$\text{因为: } L_{xx} = 75.7575, L_{xy} = 46.2425, L_{yy} = 36.8255$$

$$L_{xy} = \frac{46.2425}{36.8255} = 0.8755$$

关系。显然这一检验结果与 F 检验法的结果一致。

例 2.3 一个非线性回归的例子

行检验

表 2.3

混凝土沉降时的去除率

t/min	5	10	15	20	25	30	60
$y/\%$	38	51	58	62	64	65	67

解 (1)根据表 2.3 的数据画出散点图 (图 2-8)。

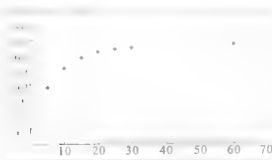


图 2-8 表 2.3 数据的散点图

(2)作变换,

令 $y' = \ln y$, $x' = \frac{1}{x}$, 则得 $y' = A + Bx' + \epsilon'$ ($\epsilon' \sim N(0, \sigma^2)$)其中, $A = \ln a$, $B = -b$ 记 $\hat{y}' = \hat{A} + \hat{B}x'$

将变换后的数据画出散点图 (图 2-9)。

回归分析。

$$x' = 0.0724, y' = 1.0421$$

$$L_{x'y} = -0.0747$$

$$L_{xx} = 0.0233$$

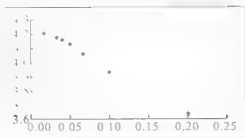


图 2-9 变换后数据的散点图

$$L_{yy'} = 0.2409$$

$$L_{xx'} = -3.2108$$

$$L_{xy'}$$

于是, 计算得: $L_{yy'} = 0.2409$, $L_{xx'} = -3.2108$, $L_{xy'} = -0.2398$ 。

$$S_M - L_W = 0.2398, S_M = 0.0011$$

$$\text{则 } F = \frac{S_M}{S_M / (n-2)} = 1.047.1872$$

查 F 分布表得: $F_{0.05}(1, 6) = 5.99$, $F_{0.01}(1, 6) = 16.24$, 因为 $1.047.1872 < 5.99$, 故在 0.05 水平上, 认为该回归方程不显著。

显著。

$$|r| = \frac{|L_{xy'}|}{\sqrt{L_{xx'} L_{yy'}}} = 0.9976$$

查相关系数表得: $r_{0.05}(6) = 0.767$, $r_{0.01}(6) = 0.989$, 查相关系数检验表得: $r_{0.05}(6) = 0.767$, $r_{0.01}(6) = 0.989$, 因为 $0.9976 > 0.989$, 故在 0.01 水平上, 认为该回归方程高度显著。列表 2-10 列出了该回归方程的统计量。

将 $y' = \ln y$, $x' = \frac{1}{x}$ 代入式(2.13)得:

$$\hat{y} = 71.8466e^{-3.2108 \cdot x}$$

【思考题2】

1. 试述变量间统计关系和函数关系的本质区别。
2. 试述回归分析与相关分析的区别与联系。
3. 一元线性回归模型有哪些基本假定?
4. 一企业排水的COD及BOD₅的结果见表2.4。

表 2.4 COD和BOD₅实测值

样品号	COD	BOD ₅	样品号	COD	BOD ₅
1	34.70	15.59	21	89.64	49.32
2	63.26	49.80	22	97.80	40.01
3	67.35	22.68	23	21.05	10.83
4	39.96	11.43	24	74.04	23.20
5	62.04	11.80	25	84.83	35.00
6	141.42	47.90	26	16.62	
7	47.84	9.56	27	61.79	33.36
8	75.23	32.36	28	88.26	28.18
9	80.61	30.40	29	138.37	52.93
10	145.05	85.08	30	122.77	51.24
11	51.80	13.61	31	52.66	17.73
12	130.07	75.02	32	92.20	25.82
13	30.17	6.02	33	145.74	56.08
14	116.20	73.76	34	117.66	45.04
15	59.00	22.08	35	69.01	26.28
16	52.86	31.68	36	79.01	24.82
17	35.54	6.90	37	81.79	38.40
18	146.51	65.64	38	98.26	44.04
19	94.75	43.32	49	125.64	58.43
20	85.53	38.26	40	142.99	73.68

- (1) 画散点图;
- (2) 判断 COD 与 BOD 之间是否大致成线性关系;
- (3) 用最小二乘法估计求回归方程;
- (4) 计算 COD 与 BOD 的决定系数;
- (5) 对回归方程作残差图, 并作分析;
- (6) 计算当 $\text{COD} = 99$ 时, BOD 的值;
- (7) 给出置信水平为 0.5 的预测区间;
7. 在一项水分渗透实验中, 得观测时间和水的重量的数据如表 2.5 所示。

表 2.5

观测时间和水的重量数据

观测时间 t/s	1	2	4	8	16	32	64
水的重量 y/g	4.22	4.02	3.85	3.59	3.44	3.02	2.59

- (1) 画出散点图;
- (2) 求曲线回归方程 $y = at^b$;
- (3) 对 $\ln y$ 与 $\ln t$ 之间的线性回归关系进行显著性检验 $\alpha = 0.05$;
6. 试用一元线性回归模型解决一个实际的环境问题。

【参考文献】

- [1] 何晓群, 现代统计分析方法与应用[M]. 北京: 中国人民大学出版社, 2003.
- [2] 卢崇飞, 多元统计推断[M]. 北京: 中国统计出版社, 1988.
- [3] 张孟威, 多元统计推断[M]. 北京: 中国统计出版社, 1984.
- [4] 陈永成, 多元统计推断[M]. 北京: 中国统计出版社, 1984.

第3章 环境多元线性回归分析

本章的主要内容是:

- 多元线性回归模型;
- 参数的最小二乘估计;
- 回归方程的显著性检验;
- 回归系数的显著性检验;
- Matlab 语言在多元线性回归中的应用;
- 环境应用。

3.1 多元线性回归模型

如下的线性关系:

式(3.1)有如下假设:

即 $\epsilon \sim N(0, \sigma^2)$ 。

得到经验回归方程:

根据多元线性关系式 $y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k + \varepsilon$, 对 $(y, x_1, x_2, \cdots,$

$x_k)$ 的第 i 次观测值 $(y_i, x_{i1}, x_{i2}, \cdots, x_{ik})$ 代入, 得

观测值:

$$(y_i, x_{i1}, x_{i2}, \cdots, x_{ik}) \quad (i=1, 2, \cdots, n; n>k+1)$$

将上述观测值代入到式(3.1)中得方程组:

$$\begin{aligned} y_1 &= b_0 + b_1x_{11} + b_2x_{12} + \cdots + b_kx_{1k} + \varepsilon_1 \\ y_2 &= b_0 + b_1x_{21} + b_2x_{22} + \cdots + b_kx_{2k} + \varepsilon_2 \\ &\vdots \end{aligned} \quad (3.3)$$

如果令:

$$Y = (y_1, y_2, \cdots, y_n)'$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

$$B = (b_0, b_1, \cdots, b_k)'$$

$$\varepsilon = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)'$$

$$\text{则可得: } Y = XB + \varepsilon \quad (3.4)$$

式(3.4)为多元线性回归模型, 在统计上, 人们常把多元线性回归模型称为多元回归分析中的问题。

3.2 参数的最小二乘估计

多元线性回归模型中, 参数 $b_0, b_1, b_2, \cdots, b_k$ 的估计, 通常采用最小二乘法。最小二乘法的基本思想是: 使残差平方和达到最小, 即

函数:

$$Q(b_0, b_1, \cdots, b_k) = \sum_{i=1}^n (y_i - b_0 - b_1x_{i1} - b_2x_{i2} - \cdots - b_kx_{ik})^2 \quad (3.5)$$

取得最小值, 即 $b_0, b_1, b_2, \cdots, b_k$ 的最佳估计。

$$\begin{aligned}\frac{\partial Q}{\partial b_0} &= -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \cdots - b_k x_{ik}) = 0 \\ \frac{\partial Q}{\partial b_1} &= -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \cdots - b_k x_{ik}) x_{i1} = 0 \\ &\vdots \\ Q &= \sum_{i=1}^n (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \cdots - b_k x_{ik}) x_{ik} = 0\end{aligned}\quad (3.6)$$

由式(3.6)可得正规方程组, 整理成矩阵形式,

$$\begin{aligned}nb_0 + \sum_{i=1}^n x_{i1}b_1 + \cdots + \sum_{i=1}^n x_{ik}b_k &= \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1}b_0 + \sum_{i=1}^n x_{i1}^2b_1 + \cdots + \sum_{i=1}^n x_{i1}x_{ik}b_k &= \sum_{i=1}^n x_{i1}y_i \\ &\vdots \\ \sum_{i=1}^n x_{ik}b_0 + \sum_{i=1}^n x_{ik}x_{i1}b_1 + \cdots + \sum_{i=1}^n x_{ik}^2b_k &= \sum_{i=1}^n x_{ik}y_i\end{aligned}\quad (3.7)$$

将式(3.7)表示成矩阵即为:

$$X'XB = X'Y \quad (\text{其中, } X' \text{ 为 } X \text{ 的转置矩阵}) \quad (3.8)$$

在式(3.8)中, X' 为 $n \times (k+1)$ 阶矩阵, X 为 $(k+1) \times n$ 阶矩阵, B 为 $(k+1) \times 1$ 阶矩阵, Y 为 $n \times 1$ 阶矩阵, 故式(3.8)可表示为

$$\hat{B} = (b_0, b_1, \dots, b_k)' = (X'X)^{-1}X'Y \quad (3.9)$$

式(3.9)即为多元线性回归系数的估计量, 式(3.9)即为多元线性回归方程。

例 3.1 为了解某地区居民的消费水平, 选取了 100 户居民进行调查, 得到如下数据:

居民编号	月收入 (元)	月消费额 (元)
1	1200	800
2	1500	1000
3	1800	1200
4	2000	1400
5	2200	1600
6	2500	1800
7	2800	2000
8	3000	2200
9	3200	2400
10	3500	2600
...
99	3800	2800
100	4000	3000

试根据上述数据, 建立多元线性回归方程, 并解释回归系数的经济意义。

表 3.1 影响国家财政的各项指标及其取值

年份	y	x_1	x_2	x	x_4	x_5	x_6
1978	1 121.1	4 237	1 397	569	96 259	1 558.6	5 076
1979	1 133.3	4 681	1 698	615	97 542	1 800.0	3 937
1980	1 080.2	5 104	1 923	767	98 705	2 116.0	4 453
1981	1 080.5	5 400	2 181	747	100 072	2 370.0	3 970
1982	1 120.0	5 811	2 483	912	101 654	2 570.0	3 313
1983	1 219.0	6 461	2 750	1 035	103 068	2 849.4	3 471
1984	1 501.9	7 617	3 214	1 263	104 357	3 376.4	3 180
1985	1 866.1	9 716	3 619	1 656	105 801	4 350.0	1 437
1986	2 260.3	11 194	4 013	2 038	107 507	4 950.0	4 711
1987	2 386.9	13 813	4 176	2 431	109 300	5 820.0	1 200
1988	2 628.0	18 224	5 865	2 967	111 026	7 410.0	5 087
1989	2 947.9	22 017	6 535	2 834	112 704	8 101.4	4 690
1990	3 244.8	23 801	7 662	3 030	114 333	8 360.1	3 817

由定性分析知, $x_1, x_2, x_3, x_4, x, x_5, x_6$ 都与变量 y 有较大的相关性。因设理论回归模型为:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x + b_6x_6 + \varepsilon_i \quad (i=1, 2, \dots, 13)$$

$$\hat{b}_0 = -460.0301; \hat{b}_1 = 0.0785; \hat{b}_2 = 0.1055; \hat{b}_3 = 0.8532;$$

$$\hat{b}_4 = -0.0011; \hat{b}_5 = -0.0078; \hat{b}_6 = 0.0045$$

则求得 y 关于 $x_1, x_2, x_3, x_4, x, x_6$ 的六元线性回归方程为:

$$\hat{y} = -460.0301 + 0.0785x_1 + 0.1055x_2 + 0.8532x_3 - 0.0011x_4 - 0.0078x + 0.0045x_6$$

完善

3.3 回归方程的显著性检验

表 3-1 多元线性回归模型中各统计量的自由度、期望值、方差、协方差

此,对于多元回归分析,一定要进行显著性检验。

多元线性回归模型的显著性检验,通常采用 F 检验。

3.3.1 拟合优度检验

多元线性回归模型的拟合优度检验,是指对多元线性回归模型的回归方程的拟合程度进行检验。通常采用 F 检验。多元线性回归模型的拟合优度检验,是指对多元线性回归模型的回归方程的拟合程度进行检验。通常采用 F 检验。

分,即:

$$S_D = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^n y_i^2 - \sum_{i=1}^n \hat{y}_i^2$$

$$= S_{\text{总}} + S_{\text{回}}$$

其中, $S_{\text{总}}$ 为总平方和, $S_{\text{回}}$ 为回归平方和, S_{D} 为残差平方和。在多元线性回归分析中, $S_{\text{总}}$ 为常数, $S_{\text{回}}$ 与 S_{D} 成正比。因此,多元线性回归模型的拟合优度检验,通常采用 F 检验。

2.2 节。

多元线性回归模型的拟合优度检验,是指对多元线性回归模型的回归方程的拟合程度进行检验。通常采用 F 检验。多元线性回归模型的拟合优度检验,是指对多元线性回归模型的回归方程的拟合程度进行检验。通常采用 F 检验。

多元线性回归模型的拟合优度检验,是指对多元线性回归模型的回归方程的拟合程度进行检验。通常采用 F 检验。多元线性回归模型的拟合优度检验,是指对多元线性回归模型的回归方程的拟合程度进行检验。通常采用 F 检验。

$$r^2 = \frac{S_{\text{回}}}{S_{\text{总}}}; r = \sqrt{\frac{S_{\text{回}}}{S_{\text{总}}}} \quad (3.10)$$

中, 共有 n 个观测值 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, 其中 n 为本国观测人数数, x_1, x_2, \dots, x_n 为观测到的自变量值, y_1, y_2, \dots, y_n 为观测到的因变量值。在多元线性回归分析中, 通常假设观测值 x_1, x_2, \dots, x_n 与 y 线性关系的显著程度, 作为一个整体的 x_1, x_2, \dots, x_n 与 y 线性关系的显著程度。

$$S_{\text{总}} = \sum_{i=1}^n (y_i - \bar{y})^2 = S_{\text{回}} + S_{\text{残}} \quad (3.1.1)$$

由此:
$$S_{\text{回}} = \sum_{i=1}^n (y_i - \hat{y})^2 = 0; \quad r^2 = \frac{S_{\text{回}}}{S_{\text{总}}} = 1 - \frac{S_{\text{残}}}{S_{\text{总}}} = 1$$

可见, 在多元线性回归分析中, 如果自变量与因变量完全线性相关, 即 n 个自变量与因变量完全线性相关, 则 $r^2 = 1$, 即 $S_{\text{回}} = S_{\text{总}}$, $S_{\text{残}} = 0$ 。

在多元线性回归分析中, 如果自变量与因变量不完全线性相关, 则 $r^2 < 1$, 即 $S_{\text{回}} < S_{\text{总}}$, $S_{\text{残}} > 0$ 。

因此, 在多元线性回归分析中, 如果自变量与因变量不完全线性相关, 则 $r^2 < 1$, 即 $S_{\text{回}} < S_{\text{总}}$, $S_{\text{残}} > 0$ 。可见, 在多元线性回归分析中, 如果自变量与因变量不完全线性相关, 则 $r^2 < 1$, 即 $S_{\text{回}} < S_{\text{总}}$, $S_{\text{残}} > 0$ 。

根据例 3.1, 给定显著性水平 $\alpha = 0.05$, 则:

$$S_{\text{回}} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = 7\,320\,924.00$$

$$S_{\text{总}} = \sum_{i=1}^n (y_i - \bar{y})^2 = 7\,439\,339.00$$

$$S_{\text{残}} = S_{\text{总}} - S_{\text{回}} = 118\,415.60$$

则样本决定系数和复相关系数分别为:

$$r^2 = \frac{S_{\text{回}}}{S_{\text{总}}} = \frac{7\,320\,924.00}{7\,439\,339.00} = 0.984\,1; \quad r = 0.992\,0$$

在显著性水平 $\alpha = 0.05$ 下, $r_{\alpha}(n-2) = r_{\alpha}(11) = 0.553$ 。

可见, 在多元线性回归分析中, 如果自变量与因变量不完全线性相关, 则 $r^2 < 1$, 即 $S_{\text{回}} < S_{\text{总}}$, $S_{\text{残}} > 0$ 。

在多元线性回归分析中, 如果自变量与因变量不完全线性相关, 则 $r^2 < 1$, 即 $S_{\text{回}} < S_{\text{总}}$, $S_{\text{残}} > 0$ 。可见, 在多元线性回归分析中, 如果自变量与因变量不完全线性相关, 则 $r^2 < 1$, 即 $S_{\text{回}} < S_{\text{总}}$, $S_{\text{残}} > 0$ 。

3.3.2 F 检验

在多元线性回归分析中, 如果自变量与因变量不完全线性相关, 则 $r^2 < 1$, 即 $S_{\text{回}} < S_{\text{总}}$, $S_{\text{残}} > 0$ 。

由式(3.10)可知, F 统计量服从自由度为 k 和 $n-k-1$ 的 F 分布。多元线性回归方程的拟合优度检验, 即检验原假设 $H_0: b_1 = b_2 = \cdots = b_k = 0$ 是否成立。如果原假设成立, 则多元线性回归方程不显著, 即多元线性回归方程不能成立。

$$H_0: b_1 = b_2 = \cdots = b_k = 0$$

如果 H_0 不成立, 则多元线性回归方程显著。因此, 多元线性回归方程的拟合优度检验, 即检验原假设 $H_0: b_1 = b_2 = \cdots = b_k = 0$ 是否成立。如果原假设成立, 则多元线性回归方程不显著, 即多元线性回归方程不能成立。

$$F = \frac{S_{\text{回}}/k}{S_{\text{残}}/(n-k-1)} \sim F(k, n-k-1) \quad (3.11)$$

由式(3.11)可知, F 统计量服从自由度为 k 和 $n-k-1$ 的 F 分布。多元线性回归方程的拟合优度检验, 即检验原假设 $H_0: b_1 = b_2 = \cdots = b_k = 0$ 是否成立。如果原假设成立, 则多元线性回归方程不显著, 即多元线性回归方程不能成立。

由式(3.11)可知, F 统计量服从自由度为 k 和 $n-k-1$ 的 F 分布。多元线性回归方程的拟合优度检验, 即检验原假设 $H_0: b_1 = b_2 = \cdots = b_k = 0$ 是否成立。如果原假设成立, 则多元线性回归方程不显著, 即多元线性回归方程不能成立。

$$F = \frac{S_{\text{回}}/k}{S_{\text{残}}/(n-k-1)} = \frac{7\,320\,924.00/6}{118\,415.60/6} = 61.824\,0$$

由式(3.11)可知, F 统计量服从自由度为 k 和 $n-k-1$ 的 F 分布。多元线性回归方程的拟合优度检验, 即检验原假设 $H_0: b_1 = b_2 = \cdots = b_k = 0$ 是否成立。如果原假设成立, 则多元线性回归方程不显著, 即多元线性回归方程不能成立。

3.4 回归系数的显著性检验

多元线性回归方程的拟合优度检验, 即检验原假设 $H_0: b_1 = b_2 = \cdots = b_k = 0$ 是否成立。如果原假设成立, 则多元线性回归方程不显著, 即多元线性回归方程不能成立。

多元线性回归方程的拟合优度检验, 即检验原假设 $H_0: b_1 = b_2 = \cdots = b_k = 0$ 是否成立。如果原假设成立, 则多元线性回归方程不显著, 即多元线性回归方程不能成立。

多元线性回归方程的拟合优度检验, 即检验原假设 $H_0: b_1 = b_2 = \cdots = b_k = 0$ 是否成立。如果原假设成立, 则多元线性回归方程不显著, 即多元线性回归方程不能成立。

容易证明, 当 H_0 成立时: $\frac{b_i}{\sqrt{C_{ii}\sigma^2}} \sim N(0, 1)$ 。

y 的影响是否显著。

根据例 3.1, 我们已经看到回归方程:

$$\hat{y} = 162.0301 + 0.0785x_1 + 0.1055x_2 + 0.8532x_3 - 0.0011x_4 - 0.3078x_5 + 0.0445x_6$$

显著影响呢? 这就需对假设 $H_{0i}: b_i = 0 (i = 1, 2, \dots, 6)$ 进行检验

利用 Matlab 软件计算, 得关于 $b_i (i = 1, 2, \dots, 6)$ 的 F 统计量 $F (i = 1, 2, \dots, 6)$, 如下:

$$F_1 = 1.0769, F_2 = 0.2974, F_3 = 2.9119, \\ F_4 = 0.0352, F_5 = 0.6597, F_6 = 0.1608$$

查 F 分布临界值表 (附表 5), 得:

$$F_{\alpha}(1, n-k-1) = F_{\alpha}(1, 6) = 5.99$$

上述 $F_i (i = 1, 2, \dots, 6) < F_{\alpha}(1, 6) = 5.99$, 即说明每一个 x_i 单独对因变

问题的定性分析选择 F 值较小的变量先剔除。

例 3.2 财政收入一项中, x_4 为人口数, $F_4 = 0.0352 < F_{\alpha}(1, 6) = 5.99$,

$$\hat{y} = 331.95 + 0.0856x_1 + 0.1063x_2 + 0.8784x_3 - 0.3428x_5 + 0.05680x_6$$

回归方程是高度显著的。

析方法详见有关参考文献。

3.5 Matlab 语言在多元回归中的应用

(1) function $[r, x]$, regress(y, x) 和 pinv(A) * y 。polyfit(x, y, n) 只能用于多项式多元线性回归, pinv(A) * y 可用于求解线性方程组。

例 3.3

表 3.2

醇类化合物拓扑指数及保留指数数据表

序号	醇	保留指数(y)			拓扑指数(A)				
		SE 30	OV 3	OV 7	1X	$^2X^*$	3X	$^4X^*$	C_{H1}
1	1-丁醇	649	673	701	2.421	0.703	0.275	0.709	
2	1-己醇	857	882	909	3.415	1.208	0.277	0.709	
3	1-庚醇	961	986	1010	3.915	1.456	0.277	0.709	
4	2-丁醇	587	608	634	2.271	0.817	0.515	0.519	
5	2-戊醇	687	710	734	2.769	0.865	0.513	0.575	
6	3-戊醇	687	709	734	2.807	1.393	0.450	0.575	
7	3-己醇	784	806	829	3.307	1.477	0.450	0.575	
8	3-庚醇	885	907	927	3.807	1.745	0.450	0.575	
9	1-庚醇	881	905	925	3.809	1.564	0.492	0.579	
10	2-甲基-2-丁醇	629	653	675	2.562	1.061	0.749	0.501	

以下代码格式为文件中真实格式)。

```

A= [1  2.424  0.703  0.275  0.709
     1  3.415  1.208  0.277  0.709
     1  3.915  1.456  0.277  0.709
     1  2.271  0.817  0.545  0.579
     1  2.769  0.865  0.543  0.575
     1  2.807  1.393  0.45  0.575
     1  3.307  1.477  0.45  0.575
     1  3.807  1.745  0.45  0.575
     1  3.809  1.564  0.452  0.579
     1  2.562  1.061  0.749  0.501];

```

```

y = [549 673 701
     877 882 909
     961 986 1010
     787 608 631
     687 710 734
     687 709 734
     781 806 829
     88  917 927
     881 905 927
     629 653 675];

```

```
b = pinv(A) * y
```

得到运行的结果为:

```
b =
```

```

-322.1033   -334.0991   319.6685
 195.3720   196.9256   192.2776
   17.8680    17.4884    19.9566
 159.0413   175.3433   175.4272
 628.9059   667.1071   701.3615

```

所以,求得的拟合方程为:

$$y_{fit} = -322.1033 - 334.0991 \chi_1 + 319.6685 \chi_2 + 195.3720 \chi_3 + 196.9256 \chi_4 + 192.2776 \chi_5 + 17.8680 \chi_6 + 17.4884 \chi_7 + 19.9566 \chi_8 + 159.0413 \chi_9 + 175.3433 \chi_{10} + 175.4272 \chi_{11} + 628.9059 \chi_{12} + 667.1071 \chi_{13} + 701.3615 \chi_{14}$$

$$y_{CH} = 319.6683 + 192.2776'X + 19.9566'X^2 + 175.4272(C^2X - X^2X^2) + 701.3615C_{CH}$$

将采取基于 Matlab 编程的方法来求解有关问题

3.6 环境应用

例 3.4

BOD 与 SS、COD 的关系, 对此进行多元线性回归分析。

表 3.3

监测结果

编号	SS	COD	BOD
1	413	45.60	13.79
2	363	37.72	12.78
3	803	70.65	26.29
4	730	81.47	23.97
5	823	90.83	28.09
6	589	58.95	18.06
7	523	50.39	17.84
8	674	61.50	22.18
9	984	107.17	34.07
10	1369	130.79	45.89

在 Matlab 中, 将表 3.3 数据输入到 M 文件, M 文件主函数为以下代码 (说明以下代码格式为文件中真实格式)。

```
A=[1 413 45.60; 1 363 37.72
1 803 70.65; 1 730 81.47
1 823 90.83; 1 589 58.95
1 523 50.39; 1 674 61.50
1 984 107.17; 1 1369 130.79];
y=[13.79;12.78;26.29;23.97;28.09;18.06;17.84;22.18;34.07;45.89];
```

$$C = A^* A$$

$$1. (C^{-1} \times Ch)^*$$

$$\begin{pmatrix} 0.000 & 0.0073 & 0.0007 \\ 0.073 & 6.0715 & 0.610 \\ 0.007 & 0.610 & 0.0618 \end{pmatrix}$$

$$\begin{pmatrix} 0.073 & 6.0715 & 0.610 \\ 0.007 & 0.610 & 0.0618 \end{pmatrix}$$

$$b = \text{pinv}(A) * y$$

$$y = A * b$$

得到运行结果为:

$$b = \begin{bmatrix} 0.5584 & 0.0293 & 0.0183 \end{bmatrix}$$

$$y = \begin{bmatrix} 13.7337 & 11.8893 & 26.9602 & 24.7152 & 27.9197 \\ 19.7301 & 17.1890 & 22.1119 & 33.4217 & 15.8330 \end{bmatrix}$$

$$\begin{bmatrix} 13.7337 & 11.8893 & 26.9602 & 24.7152 & 27.9197 \\ 19.7301 & 17.1890 & 22.1119 & 33.4217 & 15.8330 \end{bmatrix}$$

由此得到回归方程为:

$$BOD = -0.5584 + 0.0293SS + 0.0183(YH)$$

回归方程的 F 检验列于下表:

$$F = (v_1, v_2) = F(2, 7) = 9.55$$

$$r = 0.9951$$

表 3.4 回归方程的检验结果

平方来源	平方和	自由度 i	r^2	F	显著性
回归	908.1217	2	0.9951	711.8959	高度显著
残差	1.1617	7			
总合	912.861				

影响很大。

例 3.5 洛河污染分析

排污口流量 Q ; x_1 : 污水 BOD 浓度 l ; x_2 : 流过该河段所需时间 t ;

表 3.5

监测结果

编号	x_1	x_2	x_3	x_4	x_5	x_6	x_7	y
1	6.88	-0.25	27.0	674 784	11 232	477	0.083	9.35
2	6.08	2.21	27.5	477 792	11 232	193	0.083	12.3
3	2.14	-3.04	26.0	477 792	11 232	404	0.083	15.60
4	5.02	-1.75	26.0	856 224	11 232	363	0.073	5.88
5	7.89	-2.26	26.0	856 224	11 232	363	0.069	6.44
6	2.38	1.65	15.0	1 190 400	15 532	428	0.104	1.00
7	1.86	1.35	15.8	1 190 400	15 532	428	0.104	3.76
8	1.02	2.12	17.1	1 194 720	13 824	428	0.104	3.98
9	1.22	-1.92	17.5	1 194 720	13 824	428	0.104	3.68
10	0.90	-0.27	17.0	3 628 800	9 936	202	0.104	2.78
11	1.58	-0.09	17.0	3 628 800	9 936	202	0.104	1.88
12	2.78	-1.17	13.5	3 265 920	9 936	114	0.104	2.56
13	2.10	-1.30	13.5	3 265 920	9 936	114	0.104	2.72
14	2.32	-0.60	14.5	3 616 080	8 640	57.3	0.104	1.64
15	1.96	-0.60	14.5	3 616 080	8 640	57.3	0.104	2.36

以下代码格式为文件中真实格式)。

```

A = [ 1 6.88 -0.25 27.0 674784 11242 477 0.083
      1 6.08 2.21 27.5 477792 11242 193 0.083
      1 2.14 3.04 26.0 477792 11242 404 0.083
      1 5.02 0.75 26.0 856224 11242 363 0.073
      1 7.89 2.26 26.0 856224 11242 363 0.069
      1 2.38 1.65 15.0 1190400 15532 428 0.104
      1 1.86 -1.35 15.8 1190400 15532 428 0.104
      1 1.02 2.12 17.1 1194720 13824 428 0.104
      1 1.22 -1.92 17.5 1194720 13824 428 0.104
      1 0.90 0.27 17.0 3628800 9936 202 0.104
      1 1.58 0.09 17.0 3628800 9936 202 0.104
      1 2.78 1.17 13.5 3265920 9936 114 0.104
      1 2.10 1.30 13.5 3265920 9936 114 0.104
      1 2.32 0.60 14.5 3616080 8640 57.3 0.104
      1 1.96 0.60 14.5 3616080 8640 57.3 0.104];

```


$y = [9.35$
 12.30
 15.60
 5.88
 6.34
 4.00
 3.76
 3.98
 3.98
 2.78
 1.88
 2.56
 2.72
 1.64
 $2.36]$;

得到运行结果为:

 $\hat{b} =$
 9.2100
 0.3179
 -1.1490
 0.6117
 0.0000
 0.0009
 -0.0027
 202.4915
 $\hat{y} = 8.5835$
 12.7210
 13.4341
 6.9164
 6.9289
 2.3714
 2.6814
 6.2229
 6.1743
 2.3401
 1.9171
 1.8818
 2.2554
 2.2898
 2.4042
 $\hat{b} = \text{pinv}(A) * y$
 $\hat{y} = A * \hat{b}$

由此得到回归方程为:

$$\hat{y} = -9.2100 - 0.3179x_1 + 1.1490x_2 + 0.6117x_3 - 0.0000x_4 - 0.0009x_5 + 0.0027x_6 + 202.4915x_7$$

回归方程的检验列于下表:

表 3.6

回归方程的检验结果

误差来源	平方和	自由度	r^2	显著性
回归	212.8754	7	0.9070	显著
残差	21.8224	7		
总和	234.6978			

【思考题3】

1. 多元线性回归模型有哪些基本假定?

表 3.7 湖泊影响因子及其监测值

项目 年份	CO ₂ 浓度 y ($\text{mg} \cdot \text{L}^{-1}$)	农业产量 x_1 (10^4 kg)	工业总产值 $x_2/10^8$ 元	湖泊水位 x_3/m
1960	2.50	0.25	4.00	3.17
1975	2.63	0.92	21.10	3.24
1976	3.15	0.87	29.10	3.02
1977	2.52	0.60	33.00	3.24
1978	4.06	0.63	37.50	2.63
1979	3.72	0.65	42.40	2.80
1980	2.82	0.42	49.25	3.85
1981	3.31	0.40	50.00	2.97

朴国月模型

表 3.8 8 座低污染车场转量的可能因素及其测定值

	(10^4 人 \cdot km)	/辆	10^4 km	10^4 km	10^4 辆
1986	2 583	22 138	5.25	63.77	96.61
1987	1 840	23 474	5.26	66.84	111.46
1988	3 257	24 917	5.28	69.73	130.38
1989	3 034	26 304	5.32	71.69	146.43
1990	2 610	27 261	5.34	74.11	162.19
1991	2 825	27 612	5.34	76.47	183.21
1992	3 148	28 161	5.36	78.69	226.16
1993	3 479	29 645	5.38	82.21	283.98
1994	3 633	31 268	5.40	86.14	349.71
1995	3 543	32 663	5.46	91.08	417.96
1996	3 322	33 778	5.67	94.81	488.02
1997	3 544	34 346	5.76	99.75	580.36

的关系

(1) 计算出 y , x_1 , x_2 的相关系数矩阵;

(2) 求 y 与 x_1 , x_2 的元线性回归方程;

检验

表 3.9

统计数据表

时 间	耗矾率 ($\text{kg} \cdot (10 \text{ t})^{-1}$)	原水浊度	出厂水浊度
6月2日	11.2	57	0.17
6月3日	13.6	52	0.23
6月4日	13.8	47	0.24
6月5日	15.2	47	0.18
6月6日	15.0	45	0.18
6月7日	14.5	41	0.21
6月8日	15.1	42	0.21
6月9日	14.8	39	0.20
6月10日	15.2	44	0.18
6月11日	14.6	58	0.13
6月12日	15.0	72	0.13
6月13日	15.5	88	0.14
6月14日	15.3	87	0.14
6月15日	15.3	86	0.13
6月16日	14.8	112	0.14

*. 试用多元线性回归模型预测一个实际的环境问题

【参考文献】

[1] 何晓群. 现代统计分析方法与应用 [M]. 北京: 中国人民大学出版社, 2003.

刊, 1988.

第4章 环境系统聚类分析

学方法称为环境聚类分析

本章的主要内容是：

- 聚类分析概述；
- 聚类要素的数据处理；
- 距离和相似系数的计算；
- 系统聚类分析的常用方法；
- 环境应用。

4.1 聚类分析概述

相似系数法。相似系数法是指根据各变量在不同样品中的观测值,计算各变量两两之间的相似系数,再根据相似系数的大小,将变量进行分类。相似系数法分为两种:一种是系统聚类分析,另一种是快速聚类分析。系统聚类分析是指根据相似系数的大小,将变量进行分类,直到所有变量都归入一个类为止。快速聚类分析是指根据相似系数的大小,将变量进行分类,直到所有变量都归入一个类为止。

相似系数法分为两种:一种是系统聚类分析,另一种是快速聚类分析。系统聚类分析是指根据相似系数的大小,将变量进行分类,直到所有变量都归入一个类为止。快速聚类分析是指根据相似系数的大小,将变量进行分类,直到所有变量都归入一个类为止。

相似系数法分为两种:一种是系统聚类分析,另一种是快速聚类分析。系统聚类分析是指根据相似系数的大小,将变量进行分类,直到所有变量都归入一个类为止。快速聚类分析是指根据相似系数的大小,将变量进行分类,直到所有变量都归入一个类为止。

4.2 聚类要素的数据处理

聚类分析中,数据要素的处理是非常重要的。在聚类分析中,数据要素的处理是指对数据进行标准化、归一化、中心化、极差标准化等处理。数据要素的处理的目的是为了使数据要素具有可比性,从而提高聚类分析的准确性。

数据要素的处理方法有多种,常用的有:差标准化、极大值标准化、极差标准化等。

例 4.1 某市 1993 年 1 月份各站点环境指标实测值, 单位: mg/l , 各指标实测值如表 4.1 所示。

表 4.1 1993 年 1 月份各站点环境指标实测值 (单位: mg/l)

各站点	指 标				
	1	2	3	4	5
攀枝花	10.0	0.8	2.0	0.10	0.003
高 场	10.5	1.3	1.8	0.16	0.002
津 市	10.4	1.9	1.2	0.16	0.003
长 沙	8.8	2.3	1.1	0.72	0.002
中山桥	13.0	3.5	2.9	0.30	0.019
宜 城	13.4	2.3	2.4	0.02	0.005

设 x_{ij} 为第 i 个站点的第 j 个指标的实测值, $i=1, 2, \dots, m$; $j=1, 2, \dots, n$, 则除以该要素数据的总和, 即:

$$x_{ij}' = \frac{x_{ij}}{\sum_{i=1}^m x_{ij}} \quad (i=1, 2, \dots, m; j=1, 2, \dots, n) \quad (4.1)$$

$$\text{且} \quad \sum_{i=1}^m x_{ij}' = 1 \quad (j=1, 2, \dots, n)$$

```

% 例 4.1 某市 1993 年 1 月份各站点环境指标实测值, 单位: mg/l, 各指标
% 实测值如表 4.1 所示。
% 求各指标的归一化数据。
% 解: 设  $x_{ij}$  为第  $i$  个站点的第  $j$  个指标的实测值,  $i=1, 2, \dots, m$ ;  $j=1, 2, \dots, n$ ,
% 则除以该要素数据的总和, 即:
% (1) 输入: 1, 1 为行标, 2 为列标, 3 为行, 4 为列, 5 为行, 6 为列。
x=[10.0 0.8 2.0 0.10 0.003; 10.5 1.3 1.8 0.16 0.002;
    10.4 1.9 1.2 0.16 0.003; 8.8 2.3 1.1 0.72 0.002;
    13.0 3.5 2.9 0.30 0.019; 13.4 2.3 2.4 0.02 0.005];
xx=zeros(6,5);
for i=1:6
    for j=1:5
        t=sum(x,i);
        xx(i,j)=x(i,j)/t(j);
    end
end
xx

```

表 4.2 总和标准化变换结果

各站点	指 标				
	溶解氧	高锰酸盐指数	BOD	NH ₄ -N	挥发酚
紫板屯	0.151 3	0.066 1	0.175 4	0.068 5	0.088 2
高 场	0.158 9	0.107 1	0.157 9	0.109 6	0.058 8
津 市	0.157 3	0.157 0	0.105 3	0.109 6	0.088 2
长 沙	0.133 1	0.190 1	0.096 5	0.193 2	0.058 8
中上桥	0.196 7	0.280 3	0.254 4	0.205 5	0.578 8
广 城	0.202 7	0.180 1	0.210 5	0.013 7	0.117 1

(2) 标准差标准化, 即:

$$x'_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \quad (i=1, 2, \dots, m; j=1, 2, \dots, n) \quad (4.2)$$

其中, $\bar{x}_j = \frac{1}{m} \sum_{i=1}^m x_{ij}$, $s_j = \sqrt{\frac{1}{m} \sum_{i=1}^m (x_{ij} - \bar{x}_j)^2}$

且 $\sum_{i=1}^m x'_{ij} = 0$, $s_j' = \sqrt{\frac{1}{m} \sum_{i=1}^m (x'_{ij} - \bar{x}_j')^2} = 1$

处理后, 在抽样样本改变时, 它仍保持相对稳定性。

Matlab 环境下程序 (说明: 文字格式为程序中真实格式) 为:

```
x = [10.0 0.8 2.0 0.10 0.005; 10.5 1.3 1.8 0.16 0.002;
      7.1 1.9 1.2 0.16 0.005; 8.8 2.3 1.1 0.72 0.002;
      13.0 3.5 1.9 0.30 0.019; 13.4 2.3 2.4 0.02 0.005];
y = std(x,1);

xx2 = zeros(6,5);
for j = 1:5
    for i = 1:6
        xx2(i,j) = (x(i,j) - z(j))/y(j);
```

```
etal
```

```
endl
```

```
xx2
```

```
1. 计算各指标的均值和标准差, 并求出各指标的权重, 计算各指标的加权均值和标准差
```

表 4.3 标准差标准变换结果

各站点	指 标				
	溶解氧	高锰酸盐指数	BOD ₅	NH ₃ -N	挥发酚
攀枝花	-0.618 6	-1.425 9	0.158 1	-0.625 8	-0.441 1
高 场	0.314 4	0.839 9	0.158 1	-0.363 9	0.606 1
津 市	-0.375 2	0.136 7	-1.106 8	-0.363 9	0.441 1
长 沙	-1.318 7	0.332 1	-1.264 9	2.081 2	0.606 1
中洞桥	1.206 7	1.738 4	1.581 1	0.217 4	2.205 3
宜 城	1.450 1	0.332 1	0.790 6	-0.975 1	0.110 3

(3) 极大值标准化, 即:

$$x_{ij}^* = \frac{x_{ij}}{\max \{x_{ij}\}} \quad (i=1, 2, \dots, m; j=1, 2, \dots, n) \quad (4.3)$$

```
1. 计算各指标的极大值, 并求出各指标的极大值标准化后的数据
```

```
xx3=zeros(6,5); % 初始化一个6行5列的零矩阵, 用于存放极大值标准化后的数据
```

```
% 计算各指标的极大值, 并求出各指标的极大值标准化后的数据
```

```
for i=1:6 % 循环计算6个站点的极大值
```

```
    x=[10.0 0.8 2.0 0.10 0.003; 10.5 1.3 1.8 0.16 0.002;
```

```
        10.4 1.9 1.2 0.16 0.003; 8.8 2.3 1.1 0.72 0.002;
```

```
        13.0 3.5 1.9 0.30 0.019; 13.4 2.3 2.4 0.02 0.005];
```

```
    a=max(x(i,:),1);
```

```
    xx3=zeros(6,5);
```

```
    for j=1:5
```

```
        for i=1:6
```

```
            xx3(i,j)=x(i,j)/a(j);
```

```
        end
```

```
    end
```

```
xx3
```

```
% 计算各指标的极大值, 并求出各指标的极大值标准化后的数据
```


表 4.5 极差标准化变换结果

各站点	指 标				
	溶解氧	高锰酸盐指数	BOD ₅	NH ₃ -N	挥发酚
攀枝花	0.260 9	0.000 0	0.500 0	0.114 3	0.058 8
高 场	0.369 6	0.185 2	0.388 9	0.200 0	0.000 0
津 市	0.347 8	0.407 4	0.055 6	0.200 0	0.058 8
长 沙	0.000 0	0.555 6	0.000 0	1.000 0	0.000 0
中山桥	0.913 0	1.000 0	1.000 0	0.400 0	1.000 0
宜 城	1.000 0	0.555 6	0.722 2	0.000 0	0.176 5

适应某些统计方法的需要

4.3 距离和相似系数的计算

1. 在《共产党宣言》中，马克思和恩格斯指出，在旧时，人们只是从宗教、道德、法律等观念出发来解释国家，而《共产党宣言》则从经济关系出发来解释国家。在《共产党宣言》中，马克思和恩格斯指出，国家是经济关系的产物，是经济关系的集中表现。在《共产党宣言》中，马克思和恩格斯指出，国家是经济关系的产物，是经济关系的集中表现。

4.3.1 距离的计算

$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \\ 1 & 3 & 6 & 10 & 15 \\ 1 & 4 & 10 & 20 & 35 \\ 1 & 5 & 15 & 35 & 70 \end{bmatrix}$ 为 Pascal 三角阵, 即 $a_{ij} = \binom{i-1}{j-1}$, $i, j = 1, 2, \dots, 5$ 。

1. 2. 3. 4. 5. 6. 7. 8. 9. 10. 11. 12. 13. 14. 15. 16. 17. 18. 19. 20. 21. 22. 23. 24. 25. 26. 27. 28. 29. 30. 31. 32. 33. 34. 35. 36. 37. 38. 39. 40. 41. 42. 43. 44. 45. 46. 47. 48. 49. 50. 51. 52. 53. 54. 55. 56. 57. 58. 59. 60. 61. 62. 63. 64. 65. 66. 67. 68. 69. 70. 71. 72. 73. 74. 75. 76. 77. 78. 79. 80. 81. 82. 83. 84. 85. 86. 87. 88. 89. 90. 91. 92. 93. 94. 95. 96. 97. 98. 99. 100. 101. 102. 103. 104. 105. 106. 107. 108. 109. 110. 111. 112. 113. 114. 115. 116. 117. 118. 119. 120. 121. 122. 123. 124. 125. 126. 127. 128. 129. 130. 131. 132. 133. 134. 135. 136. 137. 138. 139. 140. 141. 142. 143. 144. 145. 146. 147. 148. 149. 150. 151. 152. 153. 154. 155. 156. 157. 158. 159. 160. 161. 162. 163. 164. 165. 166. 167. 168. 169. 170. 171. 172. 173. 174. 175. 176. 177. 178. 179. 180. 181. 182. 183. 184. 185. 186. 187. 188. 189. 190. 191. 192. 193. 194. 195. 196. 197. 198. 199. 200. 201. 202. 203. 204. 205. 206. 207. 208. 209. 210. 211. 212. 213. 214. 215. 216. 217. 218. 219. 220. 221. 222. 223. 224. 225. 226. 227. 228. 229. 230. 231. 232. 233. 234. 235. 236. 237. 238. 239. 240. 241. 242. 243. 244. 245. 246. 247. 248. 249. 250. 251. 252. 253. 254. 255. 256. 257. 258. 259. 260. 261. 262. 263. 264. 265. 266. 267. 268. 269. 270. 271. 272. 273. 274. 275. 276. 277. 278. 279. 280. 281. 282. 283. 284. 285. 286. 287. 288. 289. 290. 291. 292. 293. 294. 295. 296. 297. 298. 299. 300. 301. 302. 303. 304. 305. 306. 307. 308. 309. 310. 311. 312. 313. 314. 315. 316. 317. 318. 319. 320. 321. 322. 323. 324. 325. 326. 327. 328. 329. 330. 331. 332. 333. 334. 335. 336. 337. 338. 339. 340. 341. 342. 343. 344. 345. 346. 347. 348. 349. 350. 351. 352. 353. 354. 355. 356. 357. 358. 359. 360. 361. 362. 363. 364. 365. 366. 367. 368. 369. 370. 371. 372. 373. 374. 375. 376. 377. 378. 379. 380. 381. 382. 383. 384. 385. 386. 387. 388. 389. 390. 391. 392. 393. 394. 395. 396. 397. 398. 399. 400. 401. 402. 403. 404. 405. 406. 407. 408. 409. 410. 411. 412. 413. 414. 415. 416. 417. 418. 419. 420. 421. 422. 423. 424. 425. 426. 427. 428. 429. 430. 431. 432. 433. 434. 435. 436. 437. 438. 439. 440. 441. 442. 443. 444. 445. 446. 447. 448. 449. 450. 451. 452. 453. 454. 455. 456. 457. 458. 459. 460. 461. 462. 463. 464. 465. 466. 467. 468. 469. 470. 471. 472. 473. 474. 475. 476. 477. 478. 479. 480. 481. 482. 483. 484. 485. 486. 487. 488. 489. 490. 491. 492. 493. 494. 495. 496. 497. 498. 499. 500. 501. 502. 503. 504. 505. 506. 507. 508. 509. 510. 511. 512. 513. 514. 515. 516. 517. 518. 519. 520. 521. 522. 523. 524. 525. 526. 527. 528. 529. 530. 531. 532. 533. 534. 535. 536. 537. 538. 539. 540. 541. 542. 543. 544. 545. 546. 547. 548. 549. 550. 551. 552. 553. 554. 555. 556. 557. 558. 559. 560. 561. 562. 563. 564. 565. 566. 567. 568. 569. 570. 571. 572. 573. 574. 575. 576. 577. 578. 579. 580. 581. 582. 583. 584. 585. 586. 587. 588. 589. 590. 591. 592. 593. 594. 595. 596. 597. 598. 599. 600. 601. 602. 603. 604. 605. 606. 607. 608. 609. 610. 611. 612. 613. 614. 615. 616. 617. 618. 619. 620. 621. 622. 623. 624. 625. 626. 627. 628. 629. 630. 631. 632. 633. 634. 635. 636. 637. 638. 639. 640. 641. 642. 643. 644. 645. 646. 647. 648. 649. 650. 651. 652. 653. 654. 655. 656. 657. 658. 659. 660. 661. 662. 663. 664. 665. 666. 667. 668. 669. 670. 671. 672. 673. 674. 675. 676. 677. 678. 679. 680. 681. 682. 683. 684. 685. 686. 687. 688. 689. 690. 691. 692. 693. 694. 695. 696. 697. 698. 699. 700. 701. 702. 703. 704. 705. 706. 707. 708. 709. 710. 711. 712. 713. 714. 715. 716. 717. 718. 719. 720. 721. 722. 723. 724. 725. 726. 727. 728. 729. 730. 731. 732. 733. 734. 735. 736. 737. 738. 739. 740. 741. 742. 743. 744. 745. 746. 747. 748. 749. 750. 751. 752. 753. 754. 755. 756. 757. 758. 759. 760. 761. 762. 763. 764. 765. 766. 767. 768. 769. 770. 771. 772. 773. 774. 775. 776. 777. 778. 779. 780. 781. 782. 783. 784. 785. 786. 787. 788. 789. 790. 791. 792. 793. 794. 795. 796. 797. 798. 799. 800. 801. 802. 803. 804. 805. 806. 807. 808. 809. 810. 811. 812. 813. 814. 815. 816. 817. 818. 819. 820. 821. 822. 823. 824. 825. 826. 827. 828. 829. 830. 831. 832. 833. 834. 835. 836. 837. 838. 839. 840.

的变量, 记为 X_1, X_2, \dots, X_m 。其中 X_1, X_2, \dots, X_m 为第 i 个对象的第 j 个变量, $i=1, 2, \dots, N$, $j=1, 2, \dots, m$ 。这里 $N=1, 2, \dots, n$ 。

表 4.6 聚类分析数据表

| 对象 | 变 量 | | | | | |
|----|----------|----------|-----|----------|-----|----------|
| 1 | x_{11} | x_{12} | *** | x_{1j} | *** | x_{1m} |
| | x_{21} | x_{22} | *** | x_{2j} | *** | x_{2m} |
| i | x_{i1} | x_{i2} | *** | x_{ij} | *** | x_{im} |
| | x_{m1} | x_{m2} | *** | x_{mj} | *** | x_{mm} |

其中 x_{ij} 为第 i 个对象的第 j 个变量的值, $i=1, 2, \dots, N$, $j=1, 2, \dots, m$ 。

d_{ij} 应满足如下几个条件:

- (1) 非负性: $d_{ij} \geq 0$ ($i, j=1, 2, \dots, m$);
- (2) 规范性: $d_{ij}=0$ ($i=j=1, 2, \dots, m$);
- (3) 对称性: $d_{ij}=d_{ji}$ ($i, j=1, 2, \dots, m$);
- (4) 三角不等式: $d_{ij} \leq d_{ik} + d_{kj}$ ($i, j, k=1, 2, \dots, m$)。

常用的距离有:

- (1) 绝对值距离 (Kanhattan 度量或网格度量)

$$d_{ij} = \sum_{k=1}^m |x_{ik} - x_{jk}| \quad (i, j=1, 2, \dots, m) \quad (4.5)$$

- (2) 欧氏距离 (二阶 Minkowski 度量)

$$d_{ij} = \sqrt{\sum_{k=1}^m (x_{ik} - x_{jk})^2} \quad (i, j=1, 2, \dots, m) \quad (4.6)$$

欧氏距离是聚类分析中用得最广泛的距离。

- (3) 明科夫斯基 (Minkowski) 距离

$$d_{ij} = \left[\sum_{k=1}^m |x_{ik} - x_{jk}|^p \right]^{1/p} \quad (i, j=1, 2, \dots, m) \quad (4.7)$$

其中 p 为任意正实数, 当 $p=1$ 时, 即为绝对值距离; 当 $p=2$ 时, 即为欧氏距离。

(4) 切比雪夫大距离

当明科夫斯基距离 $p \rightarrow \infty$ 时,

$$d_{ij}(\infty) = \max_k |x_{ik} - x_{jk}| \quad (i, j=1, 2, \dots, m) \quad (4.8)$$

(5) Canberra 度量 (又称 U 氏距离)

$$d_{ij} = \sum_{k=1}^n \frac{|x_{ik} - x_{jk}|}{x_{ik} + x_{jk}} \quad (i, j=1, 2, \dots, m) \quad (4.9)$$

的数据。

比, 选择一种较为合理的距离进行聚类

(6) 马氏 (P. C. Mahalanobis) 距离

设 S 表示指标的协方差矩阵, 即: $S = (u_{ij})_{n \times n}$ 。

其中, $u_{ij} = \frac{1}{n-1} \sum_{k=1}^n (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j) \quad (i, j=1, 2, \dots, n)$

$$\bar{x}_i = \frac{1}{m} \sum_{k=1}^m x_{ik}, \quad \bar{x}_j = \frac{1}{m} \sum_{k=1}^m x_{jk}$$

若 S^{-1} 存在, 则两个样本之间的马氏距离为:

$$d_{ij}^2 = (\mathbf{X}_i - \mathbf{X}_j)' \mathbf{S}^{-1} (\mathbf{X}_i - \mathbf{X}_j) \quad (4.10)$$

\mathbf{X} 类似,

间相关性的影响, 可以引入斜交空间距离。

(7) 斜交空间距离



图 4-1 三维空间中,不同坐标系中用欧氏距离计算所产生的变形

斜交空间距离,其距离公式为:

$$d_{ij} = \left[\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n (x_{ai} - x_{aj})(x_{bi} - x_{bj})r_{ij} \right]^2 \quad (i, j=1, 2, \dots, m) \quad (4.11)$$

4.4.2 欧氏距离、明科夫斯基距离、斜交空间距离

欧氏距离、欧氏距离和明科夫斯基距离。

解

式为程序中真实格式)为:

```
x = [ 0.6186    1.4259    0.1581   -0.6258   -0.4411;
      -0.3114    0.8399    0.1581   -0.3639    0.6061;
       0.3752    0.1367    1.1068    0.3639    0.4411;
       1.3487    0.3321    1.2649    2.0812   -0.6061;
       1.2067    1.7384    1.5811    0.2474    2.2053;
       1.4501    0.3321    0.7906   -0.9751   -0.1103]
```

```
[m,n] = size(x);
a = zeros(m,m);
for i = 1:m
    for j = 1:m
        for k = 1:n
            a(i,j) = a(i,j) + abs(x(i,k) - x(j,k));
        end
    end
end
```

```

0.000 0
1.633 6 0.000 0
3.059 4 1.878 0 0.000 0
6.783 4 5.758 2 4.210 8 0.000 0
9.932 2 9.261 6 9.402 6 11.453 2 0.000 0
5.139 3 4.992 5 5.133 5 8.406 7 5.978 3 0.000 0
(2) 当  $p =$ 

```

为程序中真实格式) 为:

```

x=[ 0.6186 1.4259 0.1581 -0.6258 0.4411;
    0.3114 0.8399 0.1581 0.3634 0.6064;
    0.5732 0.1367 1.1068 0.3640 -0.4411;
    1.4187 0.3321 -1.2619 2.0812 -0.6064;
    1.2067 1.7384 1.5811 0.2474 2.2053;
    1.4501 0.3321 0.7906 -0.9751 -0.1103]

```

```

[m,n]=size(x);
b=zeros(m,m);
for i=1:m
    for j=1:m
        for k=1:n

```

```

        end
    end
end
sqrt(b)

```

```

0.000 0
0.794 9 0.000 0
1.841 2 1.194 0 0.000 0
3.606 1 3.105 9 2.683 0 0.000 0
4.809 9 4.501 7 4.541 0 5.279 8 0.000 0
2.828 7 2.450 9 2.763 1 4.652 5 3.085 2 0.000 0
1.450 1 0.332 1 0.790 6 -0.975 1 -0.110 3

```

4.3.2 相似系数的计算

聚类分析方法不仅可以用来对样本进行分类,而且可以对变量进行分类。在

先归为一类。相似系数定义如下:

$$(1) C_{ij} = +1 \Leftrightarrow x_i = ax_j, \quad (a \in \mathbf{R}, a \neq 0);$$

$$(2) C_{ij} \leq 1 \quad (i, j=1, 2, \dots, n);$$

相关系数, 其计算公式如下。

1. 火角余弦

图 1-2 中曲线 AB 和 CD 尽管长度不一, 但形状相似, 当长

密切的关系, 而火角余弦适合这一要求。其定义为:



图 1-2

量之火角余弦用 $\cos \theta_{ij}$ 表示, 则:

$$C_{ij} = \cos \theta_{ij} = \frac{\sum_{k=1}^m (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^m (x_{ik} - \bar{x}_i)^2} \sqrt{\sum_{k=1}^m (x_{jk} - \bar{x}_j)^2}} \quad (i, j=1, 2, \dots, m) \quad (4.12)$$

[-1, 1] 区间

矩阵:

$C_{ij} = (\cos \theta_{ij})_{n \times n}$

| | | | | | |
|----------|----------|---------|----------|----------|----------|
| 1.000 0 | 0.928 4 | 0.296 3 | -0.175 2 | 0.683 5 | 0.173 0 |
| 0.928 4 | 1.000 0 | 0.535 4 | -0.013 5 | -0.884 1 | -0.194 9 |
| 0.296 3 | 0.535 4 | 1.000 0 | 0.365 3 | -0.776 2 | 0.416 0 |
| -0.175 2 | -0.013 5 | 0.365 3 | 1.000 0 | 0.391 3 | 0.859 7 |
| -0.683 5 | 0.884 1 | 0.776 2 | 0.391 3 | 1.000 0 | 0.459 8 |
| 0.173 0 | 0.194 9 | 0.416 0 | -0.859 7 | 0.459 8 | 1.000 0 |

2 相关系数

$$r_{ij} = \frac{\sum_{k=1}^n (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^n (x_{ik} - \bar{x}_i)^2} \sqrt{\sum_{k=1}^n (x_{jk} - \bar{x}_j)^2}} \quad (i, j=1, 2, \dots, m) \quad (4.13)$$

矩阵

$R = (r_{ij})_{m \times m}$

| | | | | | | |
|--|---------|----------|----------|----------|----------|----------|
| | 1.000 0 | 0.816 8 | -0.912 8 | 0.419 1 | 0.011 6 | 0.140 6 |
| | 0.816 8 | 1.000 0 | 0.761 2 | 0.270 7 | -0.419 6 | 0.245 3 |
| | 0.912 8 | -0.761 2 | 1.000 0 | 0.470 0 | -0.111 6 | -0.253 3 |
| | 0.419 1 | -0.270 7 | 0.470 0 | 1.000 0 | -0.656 4 | 0.875 9 |
| | 0.011 6 | -0.119 6 | -0.111 6 | -0.656 4 | 1.000 0 | 0.377 3 |
| | 0.140 6 | 0.245 3 | -0.253 3 | 0.875 9 | 0.377 3 | 1.000 0 |

4.3.3 距离和相似系数选择原则

应注意相似性尺度的选择, 注意遵循下列基本选择原则:

中, 常用相关系数表示经济变量之间的亲疏程度。

空间距离, 因采用该距离处理时, 计算工作量太大。

4.4 系统聚类分析常用方法

在系统聚类分析中, 距离的定义是至关重要的。距离的定义不同, 就产生了不同的系统聚类方法。距离的定义, 一般分为两大类: 绝对距离和相对距离。绝对距离是指两个样本之间的绝对距离, 相对距离是指两个样本之间的相对距离。

绝对距离的定义如下: 设有两个样本 X_i 和 X_j , 它们的属性值为 $x_{i1}, x_{i2}, \dots, x_{in}$ 和 $x_{j1}, x_{j2}, \dots, x_{jn}$, 则它们的绝对距离 d_{ij} 定义为:

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{in} - x_{jn})^2}$$

相对距离的定义如下: 设有两个样本 X_i 和 X_j , 它们的属性值为 $x_{i1}, x_{i2}, \dots, x_{in}$ 和 $x_{j1}, x_{j2}, \dots, x_{jn}$, 则它们的相对距离 d_{ij} 定义为:

$$d_{ij} = \frac{1}{n} \sum_{k=1}^n |x_{ik} - x_{jk}|$$

绝对距离和相对距离的定义, 都是基于两个样本之间的属性值的差异来定义的。绝对距离的定义, 是基于两个样本之间的属性值的平方和来定义的, 而相对距离的定义, 则是基于两个样本之间的属性值的绝对差之和来定义的。

(1) 构造 m 个类, 每个类只包含一个样本, 记作 G_1, G_2, \dots, G_m ;

(2) 定义 m 个样本两两间的距离 $\{d_{ij}\}$, 记作 $D^{(1)} = (d_{ij}^{(1)})_{m \times m}$;

将 $D^{(1)}$ 按行、按列从小到大排序, 得到 $D^{(2)}$, 将 $D^{(2)}$ 按行、按列从小到大排序, 得到 $D^{(3)}$, 依此类推, 直到得到 $D^{(m-1)}$ 为止。

将 $D^{(m-1)}$ 按行、按列从小到大排序, 得到 $D^{(m)}$, 将 $D^{(m)}$ 按行、按列从小到大排序, 得到 $D^{(m+1)}$, 依此类推, 直到得到 $D^{(m)}$ 为止。

(5) 画聚类图:

(6) 确定临界值, 决定类的个数和类的构成。

在系统聚类分析中, 临界值的确定是非常重要的。临界值的确定, 一般分为两大类: 绝对临界值和相对临界值。绝对临界值是指两个样本之间的绝对距离, 相对临界值是指两个样本之间的相对距离。

绝对临界值的定义如下: 设有两个样本 X_i 和 X_j , 它们的属性值为 $x_{i1}, x_{i2}, \dots, x_{in}$ 和 $x_{j1}, x_{j2}, \dots, x_{jn}$, 则它们的绝对临界值 d_{ij} 定义为:

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{in} - x_{jn})^2}$$

相对临界值的定义如下: 设有两个样本 X_i 和 X_j , 它们的属性值为 $x_{i1}, x_{i2}, \dots, x_{in}$ 和 $x_{j1}, x_{j2}, \dots, x_{jn}$, 则它们的相对临界值 d_{ij} 定义为:

$$d_{ij} = \frac{1}{n} \sum_{k=1}^n |x_{ik} - x_{jk}|$$

绝对临界值和相对临界值的定义, 都是基于两个样本之间的属性值的差异来定义的。绝对临界值的定义, 是基于两个样本之间的属性值的平方和来定义的, 而相对临界值的定义, 则是基于两个样本之间的属性值的绝对差之和来定义的。

例 4.3 设有 10 个样本, 它们的属性值为 $x_{i1}, x_{i2}, \dots, x_{i10}$, 其中 $i = 1, 2, \dots, 10$ 。

过极差标准化处理 (见本章 4.2 节) 后, 如表 4.8 所示。

表 4.7 某地区九个农业区的七项指标数据

| 代号 | X_1
($\text{hm}^2 \cdot \text{人}^{-1}$) | X_2
($\text{hm}^2 \cdot \text{个}^{-1}$) | X_3
/% | X_4
/% | X_5
($\text{kg} \cdot \text{hm}^{-2}$) | X_6
($\text{kg} \cdot \text{人}^{-1}$) | 食比重 X_7
/% |
|-------|--|--|-------------|-------------|---|--|-----------------|
| G_1 | 0.294 | 1.093 | 5.63 | 113.6 | 4 510.5 | 1 036.4 | 12.20 |
| G_2 | 0.315 | 0.971 | 0.39 | 95.1 | 2 773.5 | 683.7 | 0.85 |
| G_3 | 0.123 | 0.316 | 3.28 | 148.5 | 6 934.5 | 611.1 | 6.49 |
| G_4 | 0.179 | 0.527 | 0.39 | 111.0 | 4 458.0 | 632.6 | 0.92 |
| G_5 | 0.081 | 0.212 | 72.04 | 217.8 | 12 249.0 | 791.1 | 80.38 |
| G_6 | 0.082 | 0.211 | 43.78 | 179.6 | 8 973.0 | 636.5 | 48.17 |
| G_7 | 0.075 | 0.181 | 65.15 | 194.7 | 10 689.0 | 634.3 | 80.17 |
| G_8 | 0.293 | 0.666 | 5.35 | 94.9 | 3 679.5 | 771.7 | 7.80 |
| G_9 | 0.167 | 0.414 | 2.90 | 94.8 | 4 231.5 | 574.6 | 1.17 |

表 4.8 极差标准化后的数据

| | X_1 | X_2 | X_3 | X_4 | X_5 | X_6 | X_7 |
|-------|---------|---------|---------|---------|---------|---------|---------|
| G_1 | 0.912 5 | 1.000 0 | 0.073 1 | 0.152 8 | 0.183 3 | 1.000 0 | 0.142 7 |
| G_2 | 1.000 0 | 0.866 2 | 0.000 0 | 0.002 4 | 0.000 0 | 0.236 2 | 0.000 0 |
| G_3 | 0.200 0 | 0.148 0 | 0.068 2 | 0.136 6 | 0.439 1 | 0.079 0 | 0.070 5 |
| | | | 0.000 0 | 0.131 7 | 0.177 8 | 0.125 6 | 0.000 9 |
| G_5 | 0.025 0 | 0.034 0 | 1.000 0 | 1.000 0 | 1.000 0 | 0.468 8 | 1.000 0 |
| | | | 0.605 6 | 0.689 4 | 0.651 3 | 0.134 0 | 0.595 6 |
| G_7 | 0.000 0 | 0.000 0 | 0.903 8 | 0.812 2 | 0.835 4 | 0.129 3 | 0.997 1 |
| | 0.000 0 | 0.000 0 | 0.069 2 | 0.000 8 | 0.095 6 | 0.426 8 | 0.087 1 |
| | | | 0.035 0 | 0.000 0 | 0.153 9 | 0.000 0 | 0.004 0 |

| | G | G_1 | G_2 | G_3 | G_4 | G_5 | G_6 | G_7 | G_8 | G_{10} |
|-------|---------|---------|---------|---------|---------|---------|---------|---------|-------|----------|
| G_1 | 0.000 0 | | | | | | | | | |
| G_2 | 1.534 6 | 0.000 0 | | | | | | | | |
| G_3 | 3.101 7 | 2.687 9 | 0.000 0 | | | | | | | |
| G_4 | 5.832 7 | 6.037 1 | 3.663 9 | 0.000 0 | | | | | | |
| G_5 | 4.708 7 | 4.418 2 | 1.870 4 | 1.795 8 | 0.000 0 | | | | | |
| G_6 | 5.780 0 | 5.519 5 | 2.932 1 | 0.849 8 | 1.071 3 | 0.000 0 | | | | |
| G_7 | 1.344 5 | 0.870 5 | 2.236 6 | 5.170 1 | 3.962 1 | 5.033 4 | 0.000 0 | | | |
| G_8 | 2.215 8 | 1.472 1 | 1.91 8 | 4.786 6 | 2.993 0 | 4.054 8 | 1.297 4 | 0.000 0 | | |

0.849 8, 故将 G_6 与 G_7 归并为一类, 记为 G_{11} , 即 $G_{11} = (G_6, G_7)$ 。

按此方法类推可将所有分类对象归类。

综合上述聚类过程, 可以作出最短距离聚类谱系图。



图 4-3 最短距离聚类谱系图

由图 4-3 可知, 最短距离聚类谱系图反映了分类对象之间的亲疏关系。从图中可以看出, 分类对象 G_1 和 G_2 的亲疏关系最近, 而分类对象 G_4 和 G_5 的亲疏关系最远。

本可能归不了类。

4.4.2 最远距离聚类法原理

最远距离聚类法 (Maximum Distance Method) 是一种基于距离的聚类方法。其原理是：在每次聚类过程中, 选择距离最大的两个对象进行归并, 直到所有对象都属于一个类为止。

$$d_{k,r} = \max \{d_{k,p}, d_{k,q}\} \quad (k \neq p, q) \quad (4.15)$$

在合并时, 若 $d_{ij} < d_{(i,j)k}$, 则 G_i 与 G_j 合并, 记为 $G_{(i,j)}$; 若 $d_{ij} > d_{(i,j)k}$, 则 G_i 与 G_k 合并, 记为 $G_{(i,k)}$; 若 $d_{ij} = d_{(i,j)k}$, 则 G_i 与 G_j 合并, 记为 $G_{(i,j)}$, 然后再与 G_k 合并, 记为 $G_{((i,j),k)}$ 。在合并过程中, 若遇到 $d_{ij} = d_{(i,j)k}$ 的情况, 则采用不同的公式, 而其并类步骤完全一样。

例 4.5 某地区 10 个乡 (A、B、C、D、E、F、G、H、I、J) 的土壤污染指数, 进行分类分析。

D

0, 0, 0, 0

1.534 6 0.000 0

3.101 7 2.687 9 0.000 0

2.215 8 1.472 1 1.215 8 0.000 0

— 3.832 7 6.037 4 3.663 9 4.786 6 0.000 0

1.708 7 4.448 2 1.870 4 2.993 0 1.795 8 0.000 0

2.780 0 5.519 5 2.932 1 4.054 8 0.819 8 1.071 3 0.000 0

1.344 5 0.870 5 2.236 6 1.297 4 5.170 1 3.962 1 5.033 4 0.000 0

2.632 8 1.659 0 1.191 8 0.493 3 4.855 7 3.062 1 4.123 9 1.404 8 0.000 0

根据上面的矩阵, 用最远距离聚类法聚类:

首先, 将 10 个乡 (A、B、C、D、E、F、G、H、I、J) 的土壤污染指数, 按式 (4.1.1) 计算, 得到 10 个乡 (A、B、C、D、E、F、G、H、I、J) 的土壤污染指数的距离矩阵, 记为 $D_{10 \times 10}$ 。

$G_1, G_2, G_3, G_4, G_5, G_6, G_7, G_8$ 与 G_9 之间的距离, 得:

$$d_{1,2} = \max\{d_{1,1}, d_{1,2}\} = \max\{2.215 8, 2.632 8\} = 2.632 8$$

$$d_{2,3} = \max\{d_{2,1}, d_{2,3}\} = \max\{1.472 1, 1.659 0\} = 1.659 0$$

$$d_{3,4} = \max\{d_{3,1}, d_{3,4}\} = \max\{1.215 8, 1.191 8\} = 1.215 8$$

$$d_{4,5} = \max\{d_{4,1}, d_{4,5}\} = \max\{4.786 6, 4.855 7\} = 4.855 7$$

$$d_{5,6} = \max\{d_{5,1}, d_{5,6}\} = \max\{2.993 0, 3.062 1\} = 3.062 1$$

$$d_{7,1,1} = \max\{d_{7,1}, d_{7,9}\} = \max\{4.054 8, 4.123 9\} = 4.123 9$$

$$d_{8,1,1} = \max\{d_{8,1}, d_{8,9}\} = \max\{1.297 4, 1.404 8\} = 1.404 8$$

将 G_1 与 G_2 合并, 记为 $G_{(1,2)}$; 将 G_3 与 G_4 合并, 记为 $G_{(3,4)}$; 将 G_5 与 G_6 合并, 记为 $G_{(5,6)}$; 将 G_7 与 G_8 合并, 记为 $G_{(7,8)}$; 将 $G_{(1,2)}$ 与 $G_{(3,4)}$ 合并, 记为 $G_{((1,2),(3,4))}$; 将 $G_{(5,6)}$ 与 $G_{(7,8)}$ 合并, 记为 $G_{((5,6),(7,8))}$; 将 $G_{((1,2),(3,4))}$ 与 $G_{((5,6),(7,8))}$ 合并, 记为 $G_{(((1,2),(3,4)),((5,6),(7,8)))}$; 将 $G_{(((1,2),(3,4)),((5,6),(7,8)))}$ 与 G_9 合并, 记为 $G_{((((1,2),(3,4)),((5,6),(7,8))),9)}$ 。

$$d_{ip} = \alpha_p d_{ip} + \alpha_q d_{iq}^2 + \beta d_{pi}^2 + \gamma, d_{ip}^2 = d_{iq}^2 \quad (4.16)$$

不同的值时,就形成了不同的聚类方法(表4.9)。

下面是最短距离聚类法和最远距离聚类法公式的统一:

(图4.5),最短距离为 $d_{01} = d_{a,b_1}$, 最远距离为 $d_{01} = d_{a,b_2}$ 。



图4.5 最短距离聚类法和最远距离聚类法之间的关系图

式子表示:

$$d_{ip}^2 = \alpha_p d_{ip} + \alpha_q d_{iq}^2 + \gamma + \beta d_{pi}^2, d_{ip}^2 = d_{iq}^2 \quad (4.17)$$

表4.9 聚类方法距离的统一表达式(胡永宏等, 2000)

| 方法名称 | 参 数 | | | | 距离矩阵要求 | 空间性质 |
|--------|---------------------------------------|---------------------------------------|---|----------|--------|------|
| | α_p | α_q | β | γ | | |
| 最短距离法 | 1/2 | 1/2 | 0 | 1/2 | 各种距离 | 压缩 |
| 最远距离法 | 1/2 | 1/2 | 0 | 1/2 | 各种距离 | 扩张 |
| 中间距离法 | 1/2 | 1/2 | $[-1, 4, 0]$ | 0 | 欧氏距离 | 保持 |
| 重心法 | $\frac{n_p}{n_p + n_q}$ | $\frac{n_q}{n_p + n_q}$ | $-\frac{n_p \times n_q}{(n_p + n_q)^2}$ | 0 | 欧氏距离 | 保持 |
| 类平均法 | $\frac{n_p}{n_p + n_q}$ | $\frac{n_q}{n_p + n_q}$ | 0 | 0 | 各种距离 | 保持 |
| 离差平方和法 | $\frac{n_p}{n_p + n_q}$ | $\frac{n_q}{n_p + n_q}$ | $-\frac{n_p}{n_p + n_q}$ | 0 | 欧氏距离 | 压缩 |
| 可变类 | $(1 - \frac{\beta}{2}) \frac{n_p}{n}$ | $(1 - \frac{\beta}{2}) \frac{n_q}{n}$ | < 1 | 0 | 各种距离 | 不定 |
| 可变法 | $(\frac{1 - \beta}{2})$ | $(\frac{1 - \beta}{2})$ | < 1 | 0 | 各种距离 | 扩张 |

注: n_p, n_q, n_r, n_s 分别是 G_p, G_q, G_r, G_s 的样本数目。

1. 挥发酚：以苯酚计，反映水体中低分子量酚类有机化合物的含量；

2. 总酚：以苯酚计，反映水体中酚类有机化合物的含量；

3. 4-氯酚：以4-氯酚计，反映水体中酚类有机化合物的含量；

5. 挥发酚：反映水中酚类有毒物质的含量的指标。

表 4-10 1997 年 1 月和 7 月各站点水环境监测指标实测值 单位: mg/l

| 站名 | pH | I | | | | | II | | | | |
|-----|------|------|------|-------|-------|-----|-----|-----|------|-------|---|
| | | 盐指数 | 总硬度 | 氨氮 | 亚硝酸盐 | 铁 | 盐指数 | 总硬度 | 氨氮 | 亚硝酸盐 | 铁 |
| 翠花 | 10.4 | 2.3 | 3.8 | 0.16 | 0.002 | 8.9 | 5.0 | 1.8 | 0.21 | 0.000 | |
| 望月楼 | 0.8 | 8.2 | 6.6 | 3.91 | 0.031 | 0.8 | 8.2 | 6.6 | 3.91 | 0.031 | |
| 高场 | 10.0 | 2.7 | 1.4 | 0.22 | 0.004 | 8.0 | 3.1 | 1.6 | 0.26 | 0.000 | |
| 朱沱 | 10.4 | 1.6 | 1.1 | 0.37 | 0.000 | 7.2 | 1.4 | 1.5 | 0.38 | 0.000 | |
| 寸滩 | 10.5 | 1.5 | 1.8 | 0.45 | 0.000 | 6.7 | 1.8 | 1.5 | 0.46 | 0.000 | |
| 贵阳 | 3.6 | 23.7 | 71.3 | 13.05 | 0.020 | 6.0 | 2.2 | 1.9 | 1.18 | 0.000 | |
| 张家界 | 10.2 | 2.1 | 1.4 | 0.07 | 0.000 | 8.4 | 1.4 | 1.0 | 0.37 | 0.000 | |
| 古首 | 10.2 | 2.8 | 5.0 | 0.75 | 0.003 | 7.4 | 2.4 | 1.1 | 0.47 | 0.000 | |
| 芷江 | 9.2 | 1.7 | 1.3 | 0.06 | 0.000 | 7.0 | 2.5 | 0.4 | 0.11 | 0.000 | |
| 坝上 | 10.3 | 3.6 | 1.8 | 0.00 | 0.000 | 6.6 | 2.1 | 0.8 | 0.00 | 0.000 | |
| 津市 | 10.4 | 1.8 | 0.7 | 0.35 | 0.000 | 8.5 | 2.2 | 0.9 | 0.22 | 0.000 | |
| 石门 | 10.4 | 1.7 | 0.8 | 0.36 | 0.008 | 7.1 | 2.4 | 1.2 | 0.58 | 0.000 | |
| 益阳 | 10.3 | 1.7 | 0.6 | 0.49 | 0.000 | 6.9 | 1.6 | 1.1 | 0.24 | 0.000 | |
| 湘潭 | 9.6 | 2.4 | 0.4 | 1.16 | 0.007 | 5.3 | 2.7 | 0.3 | 0.18 | 0.000 | |
| 株洲 | 5.4 | 2.8 | 1.1 | 0.28 | 0.000 | 6.2 | 2.5 | 1.1 | 0.10 | 0.006 | |
| 衡阳 | 6.8 | 2.3 | 2.9 | 0.49 | 0.007 | 7.1 | 3.3 | 1.1 | 0.20 | 0.000 | |
| 长沙 | 10.3 | 2.7 | 1.7 | 1.35 | 0.000 | 6.8 | 2.0 | 0.4 | 0.50 | 0.002 | |
| 吉安 | 9.8 | 2.0 | 2.8 | 0.00 | 0.000 | 5.6 | 2.4 | 1.4 | 0.17 | 0.000 | |
| 中山 | 12.8 | 2.1 | 2.5 | 0.10 | 0.000 | 7.3 | 0.8 | 1.2 | 0.00 | 0.027 | |
| 宜城 | 11.6 | 1.0 | 3.6 | 0.00 | 0.000 | 8.1 | 1.2 | 1.7 | 0.08 | 0.000 | |

(2) 系统聚类计算

将表 4.10 中数据按 1 月份和 7 月份分别进行标准化处理, 得到表 4.11。

将表 4.11 中数据按 1 月份和 7 月份分别进行聚类分析, 得到聚类树, 见图 4.10, D_1 、 D_2 (分别表示 1 月份和 7 月份)。

表 4.11 标准差标准化法处理后的数据

| 各
指
标 | 标准差标准化法处理后的数据 | | | | | | | | | |
|-------------|---------------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | 0.303 4 | 0.255 0 | -0.120 8 | 0.358 0 | -0.269 3 | 1.272 5 | 1.564 8 | 0.292 5 | 0.327 4 | -0.379 5 |
| 3 | 0.234 5 | 0.172 4 | -0.279 3 | 0.337 0 | -0.012 8 | 0.728 4 | 0.346 3 | 0.134 4 | -0.257 0 | 0.379 5 |
| 4 | 0.368 4 | -0.399 5 | 0.299 1 | -0.284 4 | -0.525 9 | 0.244 8 | -0.743 9 | 0.055 3 | -0.122 0 | 0.379 5 |
| 5 | 0.401 9 | 0.420 2 | -0.252 9 | 0.256 3 | -0.525 9 | -0.057 4 | 0.487 4 | 0.055 3 | 0.025 4 | -0.379 5 |
| 7 | 0.301 4 | -0.296 3 | -0.279 3 | 0.389 6 | 0.525 9 | 0.970 2 | -0.743 9 | -0.339 9 | -0.134 1 | 0.379 5 |
| 8 | 0.301 4 | 0.151 8 | -0.041 6 | 0.151 1 | -0.141 1 | 0.365 7 | -0.102 6 | 0.260 9 | -0.013 3 | 0.379 5 |
| 9 | 1.453 5 | -0.378 9 | 0.285 9 | -0.393 1 | 0.525 9 | 0.123 9 | -0.038 5 | -0.814 3 | 0.448 2 | -0.379 5 |
| 10 | 0.334 9 | 0.013 4 | -0.252 9 | 0.414 1 | -0.525 9 | -0.117 9 | -0.295 0 | 0.498 0 | -0.581 1 | 0.379 5 |
| 11 | 0.368 4 | 0.358 2 | 0.325 0 | 0.291 4 | 0.525 9 | 1.030 7 | 0.230 9 | -0.419 0 | 0.315 3 | 0.379 5 |
| 12 | 0.368 4 | -0.378 9 | 0.318 9 | -0.287 9 | 0.500 2 | 0.184 4 | 0.102 6 | -0.181 8 | 0.119 6 | -0.379 5 |
| 14 | 0.100 5 | -0.234 4 | 0.345 3 | -0.007 4 | 0.371 9 | 0.903 7 | 0.089 8 | 0.893 3 | -0.363 7 | 0.379 5 |
| 15 | 1.306 3 | 0.151 8 | -0.299 1 | 0.315 9 | -0.525 9 | -0.359 7 | -0.038 5 | 0.260 9 | -0.460 3 | 0.310 5 |
| 16 | 3.837 3 | -0.255 0 | 0.180 3 | -0.242 3 | 0.371 9 | 0.184 4 | 0.474 6 | 0.260 9 | 0.339 5 | 0.379 5 |
| 17 | 0.334 9 | 0.172 4 | -0.259 5 | 0.059 3 | -0.525 9 | 0.003 0 | 0.359 1 | 0.814 3 | 0.023 0 | 0.149 5 |
| 18 | 0.167 5 | 0.316 9 | 0.186 9 | 0.414 1 | -0.525 9 | -0.722 4 | 0.102 6 | 0.023 7 | 0.375 8 | -0.379 5 |
| 19 | 1.172 3 | 0.296 3 | 0.206 7 | 0.379 0 | 0.525 9 | 0.305 3 | 1.128 7 | 0.181 8 | -0.581 1 | 2.725 6 |
| 20 | 1.775 2 | 0.523 1 | 0.134 0 | -0.414 1 | 0.525 9 | 0.788 9 | 0.872 2 | 0.213 4 | 0.484 5 | 0.379 5 |

环境统计分析报告

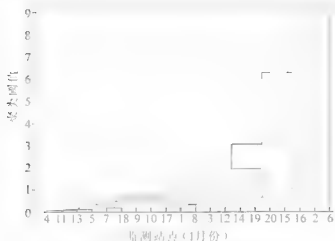


图 4-7 1997 年 1 月份 20 个水质监测站点最近距离聚类图

根据图 4-7 可知, 1 月份 20 个水质监测站点的聚类结果与图 4-6 类似, 差异性程度分为五类较合适, 即 1, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 号监测站点 1 月份水质均为Ⅲ类水, 而 2, 6 号监测站点 1 月份水质均为Ⅳ类水。

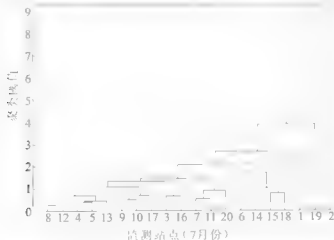


图 4-8 1997 年 7 月份 20 个水质监测站点最近距离聚类图

表 4.13 望江镇 1 月份 8 项环境指标指标 1993 ~ 2000 年, 单位: mg/l

| 年份 | 指 标 | | | | | |
|------|-----|--------|------------------|--------------------|-------|-------|
| | DO | 高锰酸盐指数 | BOD ₅ | NH ₃ -N | 挥发酚 | 铜 |
| 1993 | 7.8 | 2.9 | 6.2 | 0.47 | 0.009 | 0.000 |
| 1994 | 0.1 | 7.9 | 47.8 | 11.60 | 0.004 | 0.000 |
| 1995 | 2.1 | 9.4 | 31.8 | 3.88 | 0.004 | 0.000 |
| 1996 | 0.6 | 9.9 | 32.6 | 8.01 | 0.040 | 0.011 |
| 1997 | 1.6 | 3.9 | 14.9 | 10.50 | 0.014 | 0.000 |
| 1998 | 0.7 | 9.8 | 18.4 | 2.53 | 0.023 | 0.000 |
| 1999 | 0.9 | 10.1 | 11.0 | 2.57 | 0.016 | 0.000 |
| 2000 | 1.4 | 6.2 | 21.9 | 6.22 | 0.018 | 0.000 |

似系数, 最后选择最近距离法进行聚类分析

知距离法计算类间距离, 并对样本进行归类

表 4.14 各地区大气浓度值 单位: mg/L

| 样点 | 气 体 | | | | | |
|----|-----------------|-----------------|-------|------------------|-------------------|-----------------|
| | SO ₂ | NO ₂ | CO | PM ₁₀ | PM _{2.5} | PM ₁ |
| 1 | 0.056 | 0.084 | 0.031 | 0.038 | 0.008 1 | 0.022 0 |
| 2 | 0.049 | 0.055 | 0.100 | 0.110 | 0.022 0 | 0.007 3 |
| 3 | 0.038 | 0.130 | 0.079 | 0.170 | 0.058 0 | 0.043 0 |
| 4 | 0.034 | 0.095 | 0.058 | 0.160 | 0.200 0 | 0.029 0 |
| 5 | 0.084 | 0.066 | 0.029 | 0.320 | 0.012 0 | 0.041 0 |
| 6 | 0.064 | 0.072 | 0.100 | 0.210 | 0.028 0 | 1.380 0 |
| 7 | 0.018 | 0.089 | 0.062 | 0.260 | 0.038 0 | 0.036 0 |
| 8 | 0.069 | 0.087 | 0.027 | 0.050 | 0.089 0 | 0.021 0 |

环境统计分析

种中个属元素进行聚类,要求聚为3类

要求:(1)用标准差标准化法对土壤中重金属元素的测定结果进行处理

(2)采用欧氏距离测度10个监测站点之间样本的距离

(3)选用最远距离法计算类间距离,并对样本进行归类。

表 4.15 土壤重金属测定结果 单位: mg/kg

| 序号 | Cd | Cr | Cu | Ni | Pb | Hg | As |
|----|-------|-------|-------|-------|-------|-------|-------|
| 1 | 0.221 | 89.52 | 42.66 | 32.63 | 46.50 | 0.530 | 10.40 |
| 2 | 0.462 | 57.21 | 16.19 | 25.42 | 27.35 | 0.082 | 7.82 |
| 3 | 0.132 | 73.28 | 31.40 | 31.38 | 37.98 | 0.370 | 11.47 |
| 4 | 0.109 | 57.88 | 25.70 | 25.82 | 31.11 | 0.114 | 7.54 |
| 5 | 0.078 | 44.57 | 36.60 | 22.06 | 22.65 | 0.187 | 7.39 |
| 6 | 0.129 | 63.34 | 22.63 | 26.85 | 23.86 | 0.033 | 6.90 |
| 7 | 0.132 | 74.83 | 18.57 | 31.71 | 32.54 | 0.137 | 9.08 |
| 8 | 0.170 | 73.32 | 56.27 | 41.84 | 27.45 | 0.746 | 10.46 |
| 9 | 0.202 | 86.26 | 63.34 | 51.04 | 33.42 | 0.304 | 10.70 |
| 10 | 0.119 | 68.62 | 12.45 | 25.79 | 28.23 | 0.056 | 7.07 |

采用相似系数法对指标聚类

表 4.16 水安全指标值

| 指 标 | 陕西 | 广西 | 河南 | 江苏 | 云南 |
|---|--------|--------|--------|---------|-------|
| C ₁ 人均水资源量 10 ⁴ m ³ | 0.098 | 0.355 | 0.072 | 0.058 | 0.572 |
| C ₂ 公顷均水资源量 10 ⁴ m ³ | 0.690 | 3.615 | 0.825 | 0.855 | 3.825 |
| C ₃ 地表水利用程度 % | 13.157 | 17.619 | 18.407 | 134.752 | 5.565 |
| C ₄ 地下水利用程度 % | 28.500 | 3.013 | 41.572 | 10.739 | 0.837 |
| C ₅ 工业万元产值用水量 m ³ | 71 | 192 | 66 | 81 | 114 |
| C ₆ 农业用水综合定额 m ³ | 303 | 1176 | 197 | 478 | 593 |

续表

| 指 标 | 陕西 | 广西 | 河南 | 江苏 | 云南 |
|---|---------|---------|---------|---------|---------|
| C ₁ 人均用水量 m ³ | 220 | 650 | 220 | 600 | 340 |
| C ₂ 单位面积 COD 排放量
(t · km ⁻²) | 1.587 | 4.335 | 4.913 | 6.156 | 0.775 |
| C ₃ 工业废水处理排放达标率 % | 80.880 | 74.000 | 91.520 | 95.890 | 79.120 |
| C ₄ Ⅲ级以下水质级别占总河长
比例 % | 55.900 | 54.000 | 72.400 | 61.200 | 23.300 |
| C ₅ 侵蚀模数指数 | 1.000 | 0.264 | 0.149 | 0.094 | 0.242 |
| C ₆ 荒漠化指数 | 0.185 | 0.000 | 0.005 | 0.000 | 0.309 |
| C ₇ 森林覆盖率指数 | 0.474 | 0.497 | 0.202 | 0.074 | 0.182 |
| C ₈ 洪水受灾面积率 % | 3.852 | 5.309 | 23.920 | 1.580 | 5.840 |
| C ₉ 干旱受灾面积率 % | 32.098 | 22.687 | 29.592 | 37.537 | 3.208 |
| C ₁₀ 区域工农业产值密度/
(10 ⁸ 元 · km ⁻²) | 80.578 | 85.342 | 304.838 | 805.433 | 51.474 |
| C ₁₁ 单位面积蓄水工程库容/
(m ³ · km ⁻²) | 1.768 | 9.527 | 23.715 | 17.801 | 2.170 |
| C ₁₂ 堤防保护耕地面积率 % | 5.914 | 5.332 | 49.060 | 94.518 | 5.705 |
| C ₁₃ 人均口粮 kg | 302.108 | 340.499 | 443.118 | 417.666 | 342.304 |
| C ₁₄ 粮食单产 (kg · hm ⁻²) | 2 850 | 4 181 | 4 542 | 5 857 | 3 463 |

【参考文献】

- [1] 徐建华. 计量地理学 [M]. 北京: 高等教育出版社, 2006.
- [2] 胡永宏, 贺思辉. 综合评价方法 [M]. 北京: 科学出版社, 2000.

资源科学, 2003, 25(4): 37-42.

第 5 章 环境模糊聚类分析

模糊聚类分析是模糊数学的一个重要分支,也是模糊数学在环境统计中应用最广泛的一个领域。模糊聚类分析是模糊数学在环境统计中的一个重要应用,它主要研究如何将模糊数据按照一定的标准进行分类,从而揭示数据内在的规律和联系。模糊聚类分析在环境统计中的应用非常广泛,例如在环境质量评价、污染源识别、环境风险评估等方面都有重要的应用。模糊聚类分析的基本原理是利用模糊集合论和模糊相似关系,将数据按照一定的标准进行分类。模糊集合论是模糊数学的基础,它研究的是具有模糊性的集合。模糊相似关系是模糊数学的一个重要概念,它研究的是集合元素之间的相似程度。模糊聚类分析的基本步骤包括:确定模糊集合、计算模糊相似关系、进行模糊聚类分析等。模糊聚类分析在环境统计中的应用非常广泛,例如在环境质量评价、污染源识别、环境风险评估等方面都有重要的应用。模糊聚类分析的基本原理是利用模糊集合论和模糊相似关系,将数据按照一定的标准进行分类。模糊集合论是模糊数学的基础,它研究的是具有模糊性的集合。模糊相似关系是模糊数学的一个重要概念,它研究的是集合元素之间的相似程度。模糊聚类分析的基本步骤包括:确定模糊集合、计算模糊相似关系、进行模糊聚类分析等。

量)进行合理的分类

本章的主要内容是:

- 模糊集理论;
- 模糊相似关系和模糊等价关系;
- 模糊聚类分析步骤;
- 传递闭包法;
- 环境应用。

5.1 模糊集理论

模糊集理论是模糊数学的一个重要分支,它研究的是具有模糊性的集合。模糊集理论在环境统计中的应用非常广泛,例如在环境质量评价、污染源识别、环境风险评估等方面都有重要的应用。模糊集理论的基本原理是利用模糊集合论和模糊相似关系,将数据按照一定的标准进行分类。模糊集合论是模糊数学的基础,它研究的是具有模糊性的集合。模糊相似关系是模糊数学的一个重要概念,它研究的是集合元素之间的相似程度。模糊集理论的基本步骤包括:确定模糊集合、计算模糊相似关系、进行模糊聚类分析等。模糊集理论在环境统计中的应用非常广泛,例如在环境质量评价、污染源识别、环境风险评估等方面都有重要的应用。模糊集理论的基本原理是利用模糊集合论和模糊相似关系,将数据按照一定的标准进行分类。模糊集合论是模糊数学的基础,它研究的是具有模糊性的集合。模糊相似关系是模糊数学的一个重要概念,它研究的是集合元素之间的相似程度。模糊集理论的基本步骤包括:确定模糊集合、计算模糊相似关系、进行模糊聚类分析等。

(1) 向量表示法

$$\tilde{A} = (\tilde{A}(x_1), \tilde{A}(x_2), \dots, \tilde{A}(x_n))$$

(2) 序列表示法(序偶法)

$$\tilde{A} = \{(x, \tilde{A}(x)) | x \in X\} = \{(x_1, \tilde{A}(x_1)), (x_2, \tilde{A}(x_2)), \dots, (x_n, \tilde{A}(x_n))\}$$

(3) 分数表示法或 Zadeh 法

设论域 $X = \{x_1, x_2, \dots, x_n\}$, 则 X 上的模糊集可以写成:

$$\tilde{A} = \frac{1}{x_1} \cup \frac{1}{x_2} \cup \dots \cup \frac{1}{x_n}$$

$$\text{或 } \tilde{A} = \tilde{A}(x_1)/x_1 \cup \tilde{A}(x_2)/x_2 \cup \dots \cup \tilde{A}(x_n)/x_n$$

(4) 解析法或积分表示法

设 X 是实数域 R 上的非空子集, $\mu_{\tilde{A}}(x)$ 为模糊集 \tilde{A} 的隶属函数, 则用积分表达式表示该模糊集

当 $x \in X$ 时, $\mu_{\tilde{A}}(x)$ 为 x 属于 \tilde{A} 的程度, 当 $x \notin X$ 时 $\mu_{\tilde{A}}(x) = 0$, 此时, X 上的模糊集可以改写成:

$$\tilde{A} = \int_{x \in X} \mu_{\tilde{A}}(x)/x$$

例 5.1 设论域 $X = \{x_1, x_2, x_3, x_4, x_5\}$, x_1, x_2, x_3, x_4, x_5 分别表示“轻度污染”, “轻度污染”, “轻度污染”, “轻度污染”, “轻度污染”, 则“轻度污染”的表示方法可以写成(杨晓华等, 2005):

$$\tilde{A} = 0.0/x_1 + 0.5/x_2 + 0.7/x_3 + 0.9/x_4 + 1.0/x_5$$

$$\text{或 } \tilde{A} = \{(x_1, 0.0), (x_2, 0.5), (x_3, 0.7), (x_4, 0.9), (x_5, 1.0)\}$$

例 5.2 设论域 $X = \{x_1, x_2, x_3, x_4, x_5\}$, x_1, x_2, x_3, x_4, x_5 分别表示“轻度污染”, “轻度污染”, “轻度污染”, “轻度污染”, “轻度污染”, 则“轻度污染”的表示方法可以写成(杨晓华等, 2005):

$$\tilde{A} = 0.0/x_1 + 0.1/x_2 + 0.6/x_3 + 0.2/x_4 + 0.1/x_5$$

$$\text{或 } \tilde{A} = \{(x_1, 0.0), (x_2, 0.1), (x_3, 0.6), (x_4, 0.2), (x_5, 0.1)\}$$

例 5.3 设论域 $X = \{x_1, x_2, x_3, x_4, x_5\}$, x_1, x_2, x_3, x_4, x_5 分别表示“轻度污染”, “轻度污染”, “轻度污染”, “轻度污染”, “轻度污染”, 则“轻度污染”的表示方法可以写成(杨晓华等, 2005):

$$\tilde{A} = 0.1/x_1 + 0.2/x_2 + 0.5/x_3 + 0.2/x_4 + 0.1/x_5$$

$$\text{或 } \tilde{A} = \{(x_1, 0.1), (x_2, 0.2), (x_3, 0.5), (x_4, 0.2), (x_5, 0.1)\}$$

5.1.3 模糊集的运算

如下。

定义2 设 $\tilde{A}, \tilde{B} \in F(U)$, 若 $\forall u \in U, \tilde{B}(u) \leq \tilde{A}(u)$

$\tilde{A} = \tilde{B}$ 。显然, 包含关系具有自反性、反对称性和传递性。

定义3 设 $\tilde{A}, \tilde{B} \in F(U)$, 定义 $\tilde{A} \cup \tilde{B}, \tilde{A} \cap \tilde{B}, \tilde{A}'$, 它们分别由下列隶属函数完全刻画。

$$(\tilde{A} \cup \tilde{B})(u) = \max(\tilde{A}(u), \tilde{B}(u)) = \tilde{A}(u) \vee \tilde{B}(u)$$

例如,

$$\tilde{A} = 0.0/x_1 + 0.5/x_2 + 0.7/x_3 + 0.9/x_4 + 1.0/x_5$$

$$\tilde{B} = 0.1/x_1 + 0.2/x_2 + 0.5/x_3 + 0.2/x_4 + 0.1/x_5$$

则

$$\tilde{A} \cup \tilde{B} = 0.1/x_1 + 0.5/x_2 + 0.7/x_3 + 0.9/x_4 + 1.0/x_5$$

$$\tilde{A} \cap \tilde{B} = 0.0/x_1 + 0.2/x_2 + 0.5/x_3 + 0.2/x_4 + 0.1/x_5$$

$$\tilde{A}' = 1.0/x_1 + 0.5/x_2 + 0.3/x_3 + 0.1/x_4 + 0.0/x_5$$

模糊集的并、交、补运算具有以下性质:

(1) 交换律 $\tilde{A} \cup \tilde{B} = \tilde{B} \cup \tilde{A}$;

(3) 分配律 $\tilde{A} \cup (\tilde{B} \cap \tilde{D}) = (\tilde{A} \cup \tilde{B}) \cap (\tilde{A} \cup \tilde{D})$;

$$\tilde{A} \cap (\tilde{B} \cup \tilde{D}) = (\tilde{A} \cap \tilde{B}) \cup (\tilde{A} \cap \tilde{D});$$

(4) 吸收律 $\tilde{A} \cup (\tilde{A} \cap \tilde{B}) = \tilde{A}, \tilde{A} \cap (\tilde{A} \cup \tilde{B}) = \tilde{A}$;

(5) 幂等律 $\tilde{A} \cup \tilde{A} = \tilde{A}, \tilde{A} \cap \tilde{A} = \tilde{A}$;

(6) 对合律 $(\tilde{A}')' = \tilde{A}$;

(7) 两极律 论域 U 和空集 \emptyset 满足

$$U \cup \tilde{A} = U, U \cap \tilde{A} = \tilde{A}, \emptyset \cup \tilde{A} = \tilde{A}, \emptyset \cap \tilde{A} = \emptyset;$$

(8) 对偶律 $(\tilde{A} \cup \tilde{B})' = \tilde{A}' \cap \tilde{B}'$;

$$(\tilde{A} \cap \tilde{B})' = \tilde{A}' \cup \tilde{B}';$$

特别指出, 模糊集一般不再满足互补律, 即

$$\tilde{A} \cup \tilde{A} \neq U, \tilde{A} \cap \tilde{A}' \neq \emptyset.$$

因此, 模糊集 \tilde{A} 与 \tilde{A} 的并集不等于全集 U , 交集也不等于空集 \emptyset .

例如,

$$\tilde{A} = 0.0/x_1 + 0.5/x_2 + 0.7/x_3 + 0.9/x_4 + 1.0/x_5$$

$$\tilde{A}' = 1.0/x_1 + 0.5/x_2 + 0.3/x_3 + 0.1/x_4 + 0.0/x_5$$

则 $\tilde{A} \cup \tilde{A}' = 1.0/x_1 + 0.5/x_2 + 0.7/x_3 + 0.9/x_4 + 1.0/x_5 \neq U$

$$\tilde{A} \cap \tilde{A}' = 0.0/x_1 + 0.5/x_2 + 0.3/x_3 + 0.1/x_4 + 0.0/x_5 \neq \emptyset$$

5.1.4 模糊映射

定义 4 设 $U = \{x_1, x_2, \dots, x_n\}$ 与 $V = \{y_1, y_2, \dots, y_m\}$ 是两个有限集合, f 是从 U 到 V 的一个模糊映射, 如果存在 n 个模糊子集 $\tilde{A}_1, \tilde{A}_2, \dots, \tilde{A}_n$ 满足 $\tilde{A}_i \cap \tilde{A}_j = \emptyset$ ($i \neq j$), 且 \tilde{A}_i 的支撑集 $S(\tilde{A}_i)$ 与 V 有非空交集, 则称 f 为从 U 到 V 的一个模糊映射. 记 $f = \{\tilde{A}_1, \tilde{A}_2, \dots, \tilde{A}_n\}$, 其中 \tilde{A}_i 为 U 上的模糊子集, $i = 1, 2, \dots, n$.

它就是从 U 到 V 的一个模糊映射 f .

定义 5 设 $U = \{x_1, x_2, \dots, x_n\}$ 与 $V = \{y_1, y_2, \dots, y_m\}$ 是两个有限集合, $R = (r_{ij})_{n \times m}$ 是一个 $n \times m$ 矩阵, 其中 $r_{ij} \in [0, 1]$, 且 $\sum_{j=1}^m r_{ij} = 1$ ($i = 1, 2, \dots, n$), 则称 R 为从 U 到 V 的一个模糊矩阵.

例如,

$$R = \begin{pmatrix} 1.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 1.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.500 & 0.500 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.680 & 0.320 & 0.000 \\ 1.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 1.000 & 0.000 & 0.000 & 0.000 & 0.000 \end{pmatrix}$$

就是一个模糊矩阵.

5.2 模糊关系

定义 6 设 $U = \{x_1, x_2, \dots, x_n\}$ 与 $V = \{y_1, y_2, \dots, y_m\}$ 是两个有限集合, $R = \{r_{ij} \mid i = 1, 2, \dots, n; j = 1, 2, \dots, m\}$ 为 U 到 V 的一个模糊关系, 记为 $U \xrightarrow{R} V$.

设 \tilde{R} 为集合 $U = \{u_1, u_2, \dots, u_n\}$ 到 $V = \{v_1, v_2, \dots, v_m\}$ 的一个模糊关系, $\gamma = \{u_1, u_2, \dots, u_n\} \in U, \delta = \{v_1, v_2, \dots, v_m\} \in V$, 则模糊关系 \tilde{R} 可用如下的模糊矩阵 R 来表示:

$$R = (r_{ij})_{n \times m}$$

其中, $r_{ij} = \mu_{\tilde{R}}(u_i, v_j) \in [0, 1]$ 。

例如, 设 $U = \{x_1, x_2, x_3\}$ 为某类植物, $V = \{y_1, y_2, y_3, y_4\}$ 为另一类植物, 且 $f: U \rightarrow V$ 为模糊函数, 其模糊关系 \tilde{R} 可用模糊矩阵 R 表示如下:

$$R = \begin{pmatrix} 0.1 & 0.2 & 0.3 & 0.6 \\ 0.7 & 0.4 & 0.6 & 0.9 \\ 0.5 & 0.8 & 0.9 & 0.7 \end{pmatrix}$$

由 f 可以确定一个模糊关系矩阵 R :

$$R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nm} \end{pmatrix}$$

例如, 设 $U = \{x_1, x_2, x_3\}$ 为某类植物, $V = \{y_1, y_2, y_3, y_4\}$ 为另一类植物, 且 $f: U \rightarrow V$ 为模糊函数, 其模糊关系 \tilde{R} 可用模糊矩阵 R 表示如下:

$$R = \begin{pmatrix} 0.1 & 0.2 & 0.3 & 0.6 \\ 0.7 & 0.4 & 0.6 & 0.9 \\ 0.5 & 0.8 & 0.9 & 0.7 \end{pmatrix}$$

其中, r_{ij} 表示 x_i 与 y_j 的相似程度, 例如, $r_{11} = 0.1$ 表示 x_1 与 y_1 的相似程度为 0.1, 即基本上不相像。

模糊矩阵 R 满足 $R^T = R$ 时, 称 R 为模糊等价关系, 即有 $r_{ij} = r_{ji}$, 且 $r_{ii} = 1$ 对所有 i 都成立。

定义 7 设 $R = (r_{ij})_{n \times n}$, $Q = (q_{ij})_{n \times n}$, $r_{ij}, q_{ij} \in [0, 1]$, 且 $R \cap Q = (r_{ij} \wedge q_{ij})_{n \times n}$, $R \cup Q = (r_{ij} \vee q_{ij})_{n \times n}$ 。

若 $U = V$, 而 $\tilde{R} \in F(U \times U)$, 则记

$$\tilde{R} \circ \tilde{R} = \tilde{R}^2, \tilde{R}^3 = \tilde{R}^2 \circ \tilde{R}, \dots, \tilde{R}^n = \tilde{R}^{n-1} \circ \tilde{R}, \dots$$

式中, “ \vee ” “ \wedge ” 分别为 “取大” “取小” 运算。

5.3 模糊等价关系

模糊等价关系的定义如下:

定义 8 设 $R = (r_{ij})_{n \times n}$ 为模糊矩阵, $r_{ij} \in [0, 1]$, $i, j = 1, 2, \dots, n$. 称 R 为模糊相似矩阵 $R = (r_{ij})_{n \times n}$, 若满足:

- (1) 自反性: $r_{ii} = 1$;
- (2) 对称性: $r_{ij} = r_{ji}$;
- (3) 传递性: $R \cdot R \subseteq R$.

若 R 同时满足自反性和对称性, 则称 R 为模糊等价矩阵. 模糊相似矩阵 R 的自反性和对称性则为相似关系.

例如, 这甲模糊相似矩阵 R 平方定义为:

$$R \cdot R = (S_{ij})_{n \times n}$$

其中 $S_{ij} = \max_{k=1, 2, \dots, n} \{ \min(r_{ik}, r_{kj}) \}$, $i, j = 1, 2, \dots, n$. 故

$$R = \begin{bmatrix} 1.0 & 0.5 & 0.5 \\ 0.8 & 0.5 & 1.0 \end{bmatrix}$$

显然 R 具有自反性, 由

$$R \cdot R = \begin{bmatrix} 1.0 & 0.5 & 0.8 \\ 0.8 & 0.5 & 1.0 \end{bmatrix} \subseteq \begin{bmatrix} 1.0 & 0.5 & 0.8 \\ 0.8 & 0.5 & 1.0 \end{bmatrix} = R$$

可见 R 也具有传递性, 故 R 是模糊等价矩阵.

定义 9 λ 截矩阵 R_λ : 设矩阵 $R = (r_{ij})_{n \times m}$, 即:

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nm} \end{bmatrix}$$

记 $R_\lambda = (r_{ij}(\lambda))_{n \times m}$.

若

$$r_{ij}(\lambda) = \begin{cases} 1 & (r_{ij} \geq \lambda) \\ 0 & (r_{ij} < \lambda) \end{cases}$$

则称 R_λ 为 R 的 λ 截矩阵 R_λ .

例如, 设 $R = \begin{bmatrix} 0.8 & 0.5 & 0.5 \\ 0.8 & 0.5 & 1.0 \end{bmatrix}$, 取 $\lambda = 0.5$, 则 $R_{0.5} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$.

定理 1 设 $R = (r_{ij})_{n \times m}$ 为模糊相似矩阵, $\lambda \in [0, 1]$, 且 $R_\lambda \subseteq R$, 则 R_λ 也是模糊相似矩阵. 若 R 是模糊等价矩阵, 则 R_λ 也是模糊等价矩阵. 若 R 是模糊相似矩阵, 则 R_λ 也是模糊相似矩阵. 若 R 是模糊等价矩阵, 则 R_λ 也是模糊等价矩阵. 若 R 是模糊相似矩阵, 则 R_λ 也是模糊相似矩阵. 若 R 是模糊等价矩阵, 则 R_λ 也是模糊等价矩阵.

为 $t(R)$ 。

定理2 设 $t(R)$ 为 R 的 λ 截矩阵, $\lambda \in [0, 1]$, 则 $t(R)$ 也是等价关系。

证明 首先, $t(R)$ 为 R 的 λ 截矩阵, 故 $t(R)$ 为 R 的子集, 又 R 为等价关系, 故 $t(R)$ 也是等价关系。

定理3 设 $t(R)$ 为 R 的 λ 截矩阵, 则 $t(R)$ 为 R 的 λ 截矩阵, 称 R_λ 分类法是 R 分类法的细化。

证明 首先, $t(R)$ 为 R 的 λ 截矩阵, 故 $t(R)$ 为 R 的子集, 又 R 为等价关系, 故 $t(R)$ 也是等价关系。又 $t(R)$ 为 R 的 λ 截矩阵, 故 $t(R)$ 为 R 的 λ 截矩阵, 故 $t(R)$ 为 R 的 λ 截矩阵。由定理2和定理3加细分类。

5.4 模糊聚类分析步骤

模糊聚类分析步骤如下: 1. 数据标准化; 2. 模糊相似矩阵的构造; 3. 模糊聚类分析。

5.4.1 数据标准化

1. 数据标准化的作用

在模糊聚类分析中, 数据标准化是一个非常重要的步骤。数据标准化是指将原始数据中的各个元素, 按照一定的标准, 将其值归一化到 $[0, 1]$ 的范围内。

2. 数据变换

设 $X = \{x_{ij} | i = 1, 2, \dots, n; j = 1, 2, \dots, m\}$ 为原始数据矩阵, 其中 x_{ij} 表示第 i 个元素有:

$$x_{ij} = \{x_{i1}, x_{i2}, \dots, x_{im}\} \quad (i = 1, 2, \dots, n)$$

这时原始数据矩阵为:

$$\begin{matrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{matrix}$$

(1) 标准差变换

$$x'_{ik} = \frac{x_{ik} - \bar{x}_k}{s_k} \quad (i=1, 2, \dots, n; k=1, 2, \dots, m)$$

其中, $\bar{x}_k = \frac{1}{n} \sum_{i=1}^n x_{ik}, \quad s_k = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_{ik} - \bar{x}_k)^2}$

可见, x'_{ik} 是 x_{ik} 的无量纲化, 且 $x'_{ik} \in [-1, 1]$, 但 x'_{ik} 不一定在 $[0, 1]$ 区间上。

(2) 极差变换

$$x''_{ik} = \frac{x_{ik} - \min_{1 \leq j \leq n} \{x_{jk}\}}{\max_{1 \leq j \leq n} \{x_{jk}\} - \min_{1 \leq j \leq n} \{x_{jk}\}} \quad (k=1, 2, \dots, m)$$

可见, x''_{ik} 是 x_{ik} 的无量纲化, 且 $x''_{ik} \in [0, 1]$, 故称 x''_{ik} 为 x_{ik} 的极差变换。

5.4.2 模糊相似矩阵的建立

模糊相似矩阵的建立, 是模糊聚类分析中可相似矩阵的建立, 即 $r_{ij} \in [0, 1] \quad (i, j=1, 2, \dots, n)$ 。

设 x_1, x_2, \dots, x_n 为 n 个待分类对象, 且 $x_i = (x_{i1}, x_{i2}, \dots, x_{im})^T$ 为 $n \times m$ 矩阵 $R = (R_{ij})$ 的行向量, 其中 $R_{ij} = x_{ij} \quad (i=1, 2, \dots, n; j=1, 2, \dots, m)$ 。

下面介绍建立模糊相似矩阵 r_{ij} 的常用方法, 并给出 r_{ij} 的性质。

1. 相似系数法

(1) 数量积法

$$r_{ij} = \begin{cases} 1 & (i=j) \\ \frac{1}{M} \sum_{k=1}^m x_{ik} x_{jk} & (i \neq j) \end{cases}$$

$$M = \max_{(i,j)} \left(\sum_{k=1}^m x_{ik} x_{jk} \right)$$

可见, r_{ij} 是 x_i 与 x_j 的相似系数, 且 $r_{ij} \in [0, 1]$, 故称 r_{ij} 为数量积法。

1

则 $r'_{ij} \in [0, 1]$ 。

类,是最细的分类。

(2) 当 $\lambda \geq 0.80$ 时

此时, $R_{\lambda} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$

可以看出共分 4 类: $\{u_1, u_3\}, \{u_2\}, \{u_4\}, \{u_5\}$ 。

(3) 当 $\lambda \geq 0.50$ 时

$$R_{\lambda} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

可以看出共分 3 类: $\{u_1, u_2, u_3\}, \{u_4\}, \{u_5\}$ 。

(4) 当 $\lambda \geq 0.45$ 时

$$R_{\lambda} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 \end{bmatrix}$$

可以看出共分 2 类: $\{u_1, u_2, u_3, u_5\}, \{u_4\}$ 。

(5) 当 $\lambda \geq 0.40$ 时

此时, $R_{\lambda} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 \end{bmatrix}$

2. 布尔矩阵法

设 $R = (r_{ij})_{n \times n}$, $i, j = 1, 2, \dots, n$, $r_{ij} \in \{0, 1\}$, 其中, r_{ij} 为 1 表示 u_i 与 u_j 属于同一类, 否则为 0。

由 R 可求出 R 的布尔幂 R^2, R^3, \dots, R^k 。

若 $R^k = R^{k+1}$, 则 R^k 为 R 的布尔闭包。

若 R 为 n 阶布尔矩阵, 则 $R \in \{0, 1\}^{n \times n}$, 若 R 没有 n 阶布尔闭包, 则 R 不是

矩阵。

在讨论矩阵法分类时所得的模糊相似矩阵为:

$$R = \begin{pmatrix} 1.00 & 0.50 & 0.80 & 0.40 & 0.45 \\ 0.50 & 1.00 & 0.30 & 0.40 & 0.45 \\ 0.80 & 0.50 & 1.00 & 0.10 & 0.45 \\ 0.40 & 0.40 & 0.10 & 1.00 & 0.40 \\ 0.45 & 0.45 & 0.40 & 0.40 & 1.00 \end{pmatrix}$$

其最大树见图 5-1。



图 5-1 模糊聚类图

u_1, u_2 ;

(2) 取 $\lambda = 0.80$, 这时分 4 类: $\{u_1, u_2\}, \{u_3\}, \{u_4\}, \{u_5\}$;

(3) 取 $\lambda = 0.50$, 这时分 3 类: $\{u_1, u_2, u_3\}, \{u_4\}, \{u_5\}$;

(4) 取 $\lambda = 0.45$, 这时分 2 类: $\{u_1, u_2, u_3, u_5\}, \{u_4\}$;

(5) 取 $\lambda = 0.40$, 这时为 1 类: $\{u_1, u_2, u_3, u_4, u_5\}$ 。

3. 编网法

设 $U = \{u_1, u_2, u_3, u_4, u_5\}$ 为“ U ”的 5 个元素, U 中任意两元素 u_i, u_j 的

建立 λ 截矩阵:

$$R_{\lambda} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

又 $\lambda = 0.40$ 时 $R_{\lambda} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$, 在 R_{λ} 中 $r_{12} = r_{21} = 1$

故 u_1, u_2 属于同一类, 记为 $U_1 = \{u_1, u_2\}$, 又 $r_{34} = r_{43} = 1$, 故有 u_3, u_4 属于同一类, 记为 $U_2 = \{u_3, u_4\}$

u_5 独自一类, 记为 $U_3 = \{u_5\}$, 故 $U = U_1 \cup U_2 \cup U_3 = \{u_1, u_2, u_3, u_4, u_5\}$

"

"

图 5-2 由 $\lambda=0.80$ 截矩阵改造后的编网图

系,由此获得的分类数是 2: $\{u_1, u_3\}$, $\{u_2, u_4, u_5\}$ 。

的结论基本一致。

5.4.4 分类的 F 检验

表 5.1。

表 5.1

原始数据表

| 样本 | 指 标 | | | | | |
|-----|----------|----------|-----|----------|-----|----------|
| | 1 | 2 | ... | k | ... | m |
| u | x_{11} | x_{12} | ... | x_{1k} | ... | x_{1m} |
| | x_{21} | x_{22} | ... | x_{2k} | ... | x_{2m} |

总体样本的中心向量为:

$$\bar{u} = (\bar{u}_1, \bar{u}_2, \dots, \bar{u}_k, \dots, \bar{u}_m)$$

其中:

$$\bar{u}_k = \frac{1}{n} \sum_{j=1}^r \sum_{i=1}^{n_j} x_{ij}^{(k)} = \frac{1}{n} \sum_{j=1}^r \sum_{i=1}^{n_j} (x_{ij}^{(1)} + 1, x_{ij}^{(2)} + 1, \dots, x_{ij}^{(k)} + 1, \dots, x_{ij}^{(m)} + 1) \\ = u_1^{(j)}, \dots, u_m^{(j)}。第j类聚类中心向量为:$$

$$u^{(j)} = (u_1^{(j)}, u_2^{(j)}, \dots, u_m^{(j)})$$

式中:

$$u_k^{(j)} = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij}^{(k)} \quad (k=1, 2, \dots, m)$$

作F统计量:

$$F = \frac{\sum_{j=1}^r \|u^{(j)} - \bar{u}\|^2 / (r-1)}{\sum_{j=1}^r \sum_{i=1}^{n_j} \|u_i^{(j)} - u^{(j)}\|^2 / (n-r)}$$

其中:

$$\|u^{(j)} - \bar{u}\| = \sqrt{\sum_k (u_k^{(j)} - \bar{u}_k)^2}$$

为 $u^{(j)}$ 与 \bar{u} 的距离。

$$\|u_i^{(j)} - u^{(j)}\|$$

为第j类中样本 $u_i^{(j)}$ 与中心 $u^{(j)}$ 的距离。

若 $F > F_{\alpha}(r-1, n-r)$, 则说明类与类之间的差异大, 分类明显。

表 5.2 模糊聚类的F检验表

| 分类数 | 检验统计量F值 | 临界值 F_{α} |
|------|---|------------------------|
| | $\sum_{j=1}^r \ u^{(j)} - \bar{u}\ ^2 / (r-1)$ | $F_{\alpha}(r-1, n-r)$ |
| | $\sum_{j=1}^r \sum_{i=1}^{n_j} \ u_i^{(j)} - u^{(j)}\ ^2 / (n-r)$ | |
| 统计推断 | 拒绝分类数 r | $F < F_{\alpha}$ |
| | 接受分类数 r | $F > F_{\alpha}$ |

表 5.3

标准化的数据

| 年份 | 指 标 | | | | | |
|------|---------|---------|---------|---------|---------|---------|
| | 100 | 1000000 | 1000 | 100000 | 1000000 | 100 |
| 1993 | 1.000 0 | 0.000 0 | 0.000 0 | 0.000 0 | 0.138 9 | 0.000 0 |
| 1994 | 0.000 0 | 0.694 4 | 1.000 0 | 1.000 0 | 0.000 0 | 0.000 0 |
| 1995 | 0.259 7 | 0.902 8 | 0.615 4 | 0.306 4 | 0.000 0 | 0.000 0 |
| 1996 | 0.064 9 | 0.972 2 | 0.634 6 | 0.677 4 | 1.000 0 | 1.000 0 |
| 1997 | 0.194 8 | 0.138 9 | 0.209 1 | 0.901 2 | 0.277 8 | 0.000 0 |
| 1998 | 0.077 9 | 0.958 3 | 0.293 3 | 0.185 1 | 0.527 8 | 0.000 0 |
| 1999 | 0.103 9 | 1.000 0 | 0.836 5 | 0.188 7 | 0.333 3 | 0.000 0 |
| 2000 | 0.168 8 | 0.458 3 | 0.449 5 | 0.516 6 | 0.388 9 | 0.000 0 |

(3) 建立模糊相似矩阵

设 x_1, x_2, \dots, x_n 为 n 个对象, r_{ij} 为对象 x_i 与对象 x_j 的相似程度, 记为 $r_{ij}(i, j=1, 2, \dots, n)$ 。

采用夹角余弦法计算:

$$r_{ij} = \frac{x_i \cdot x_j}{\sqrt{x_i^2} \cdot \sqrt{x_j^2}}$$

式中, x_i 为 n 维向量, x_i^2 为 n 维向量 x_i 的模, $x_i \cdot x_j$ 为向量 x_i 与向量 x_j 的内积。

于是, r_{ij} 为 n 阶模糊相似矩阵, 即 $r_{ij} \in [0, 1]$ 。由模糊相似矩阵 r_{ij} 可求得模糊等价矩阵 R , 即由 r_{ij} 中每一元素 r_{ij} 求其平方根 $\sqrt{r_{ij}}$, 所得 $\sqrt{r_{ij}}$ 元素, 即为模糊等价矩阵 R , 故知 R 为 n 阶模糊等价的一个对称矩阵。

$\{u_8\}$ 。

- 当 $\lambda=0.879\ 7$ 时,分5类: $\{u_1\}$, $\{u_2, u_8\}$, $\{u_3, u_5\}$, $\{u_6, u_7\}$, $\{u_4\}$ 。
- 当 $\lambda=0.854\ 4$ 时,分4类: $\{u_1\}$, $\{u_2, u_3, u_6, u_7\}$, $\{u_3, u_5\}$, $\{u_4\}$ 。
- 当 $\lambda=0.828\ 6$ 时,分3类: $\{u_1\}$, $\{u_2, u_3, u_4, u_5, u_6, u_7, u_8\}$, $\{u_4\}$ 。
- 当 $\lambda=0.818\ 5$ 时,分2类: $\{u_1\}$, $\{u_2, u_3, u_4, u_5, u_6, u_7, u_8\}$ 。
- 当 $\lambda=0.238\ 2$ 时,分1类: $\{u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8\}$ 。

例 5.5 某市环保局为评价道路大气环境质量,选取了CO、SO₂、NO_x、PM₁₀和TSP 5个评价指标。

分别是CO、SO₂、NO_x、PM₁₀和TSP,构造了4个评价等级标准,分别为:优、良、中、差(参考《环境空气质量标准》(GB 3095-1996)和《环境空气质量标准》(GB 3095-2012)标准,2004)。

表 5.4 道路大气环境质量评价标准等级值 单位: mg/m³

| 等级 | 评价指标 | | | | |
|------------|------|-----------------|-----------------|------------------|------|
| | CO | SO ₂ | NO _x | PM ₁₀ | TSP |
| 优(u_1) | 3.0 | 0.05 | 0.05 | 0.05 | 0.12 |
| 良(u_2) | 3.5 | 0.10 | 0.08 | 0.10 | 0.20 |
| 中(u_3) | 4.0 | 0.15 | 0.10 | 0.15 | 0.30 |
| 差(u_4) | 6.0 | 0.25 | 0.15 | 0.25 | 0.50 |

表 5.5 道路大气环境质量实测值 单位: mg/m³

| | 评价指标 | | | | |
|--------------|------|-----------------|-----------------|------------------|------|
| | CO | SO ₂ | NO _x | PM ₁₀ | TSP |
| 公路a(u_1) | 2.6 | 0.06 | 0.05 | 0.05 | 0.10 |
| 公路b(u_2) | 4.7 | 0.16 | 0.09 | 0.08 | 0.31 |
| 公路c(u_3) | 3.7 | 0.14 | 0.09 | 0.09 | 0.21 |
| 公路d(u_4) | 3.0 | 0.13 | 0.07 | 0.08 | 0.18 |

根据表5.4和表5.5,可以计算出各评价指标的模糊子集,进而求出道路大气环境质量的模糊聚类分析结果。

表 5.6

标准化数据

| | \bar{x}_i | \bar{x}_{ii} | \bar{x}_{ij} | \bar{x}_{ji} | \bar{x}_{jj} |
|------|-------------|----------------|----------------|----------------|----------------|
| 优 | 0.117 6 | 0.000 0 | 0.000 0 | 0.000 0 | 0.050 0 |
| 良 | 0.264 7 | 0.250 0 | 0.300 0 | 0.250 0 | 0.250 0 |
| 中 | 0.411 8 | 0.500 0 | 0.500 0 | 0.500 0 | 0.500 0 |
| 差 | 1.000 0 | 1.000 0 | 1.000 0 | 1.000 0 | 1.000 0 |
| 公路 a | 0.000 0 | 0.050 0 | 0.000 0 | 0.000 0 | 0.000 0 |
| 公路 b | 0.617 6 | 0.550 0 | 0.400 0 | 0.150 0 | 0.525 0 |
| 公路 c | 0.323 5 | 0.450 0 | 0.400 0 | 0.200 0 | 0.275 0 |
| 公路 d | 0.117 6 | 0.400 0 | 0.200 0 | 0.150 0 | 0.200 0 |

步骤 3 计算模糊等价矩阵 R ，由模糊合成法得：

$$\begin{aligned}
 &1.000\ 0\ 0.579\ 1\ 0.531\ 3\ 0.586\ 5\ 0.000\ 0\ 0.724\ 0\ 0.530\ 9\ 0.354\ 8 \\
 &0.579\ 1\ 1.000\ 0\ 0.994\ 4\ 0.997\ 3\ 0.424\ 1\ 0.936\ 0\ 0.970\ 7\ 0.899\ 1 \\
 &0.531\ 3\ 0.994\ 4\ 1.000\ 0\ 0.997\ 3\ 0.462\ 3\ 0.923\ 1\ 0.963\ 8\ 0.920\ 7
 \end{aligned}$$

R

$$\begin{aligned}
 &0.000\ 0\ 0.424\ 1\ 0.462\ 3\ 0.447\ 2\ 1.000\ 0\ 0.514\ 6\ 0.589\ 5\ 0.760\ 9 \\
 &0.724\ 0\ 0.936\ 0\ 0.923\ 1\ 0.938\ 5\ 0.514\ 6\ 1.000\ 0\ 0.958\ 1\ 0.890\ 3 \\
 &0.530\ 9\ 0.970\ 7\ 0.963\ 8\ 0.965\ 7\ 0.589\ 5\ 0.958\ 1\ 1.000\ 0\ 0.954\ 5 \\
 &0.354\ 8\ 0.899\ 1\ 0.920\ 7\ 0.908\ 3\ 0.760\ 9\ 0.890\ 3\ 0.954\ 5\ 1.000\ 0
 \end{aligned}$$

步骤 4 计算模糊等价矩阵 R^* ，由模糊合成法得：

$R^* = R^4$ ，故 R^* 就是所求的模糊等价矩阵。

$$\begin{aligned}
 &1.000\ 0\ 0.724\ 0\ 0.724\ 0\ 0.724\ 0\ 0.724\ 0\ 0.724\ 0\ 0.724\ 0\ 0.724\ 0 \\
 &0.724\ 0\ 1.000\ 0\ 0.997\ 3\ 0.997\ 3\ 0.760\ 9\ 0.958\ 1\ 0.970\ 7\ 0.954\ 5 \\
 &0.724\ 0\ 0.997\ 3\ 1.000\ 0\ 0.997\ 3\ 0.760\ 9\ 0.958\ 1\ 0.970\ 7\ 0.954\ 5 \\
 &0.724\ 0\ 0.997\ 3\ 0.997\ 3\ 1.000\ 0\ 0.760\ 9\ 0.958\ 1\ 0.970\ 7\ 0.954\ 5
 \end{aligned}$$

R

$$\begin{aligned}
 &0.724\ 0\ 0.958\ 1\ 0.958\ 1\ 0.958\ 1\ 0.760\ 9\ 1.000\ 0\ 0.958\ 1\ 0.954\ 5 \\
 &0.724\ 0\ 0.970\ 7\ 0.970\ 7\ 0.970\ 7\ 0.760\ 9\ 0.958\ 1\ 1.000\ 0\ 0.954\ 5 \\
 &0.724\ 0\ 0.954\ 5\ 0.954\ 5\ 0.954\ 5\ 0.760\ 9\ 0.954\ 5\ 0.954\ 5\ 1.000\ 0
 \end{aligned}$$

在 R 中, 元素 x_i 与 x_j 的相似程度, 记为 r_{ij} , 根据相似矩阵 R 的构造方法, 可以求出 r_{ij} 的值, 从而得到相似矩阵 R 。根据相似矩阵 R 的不同, 可以得到不同的聚类结果, 可以用动态聚类图来表示(图 5-4)。



图 5-4 动态聚类图

由此可以得出结论:

- 当 $\lambda=0.9973$ 时, 分 7 类: $\{u_1\}$, $\{u_2\}$, $\{u_3, u_4\}$, $\{u_5\}$, $\{u_6\}$, $\{u_7\}$, $\{u_8\}$ 。
- 当 $\lambda=0.9707$ 时, 分 5 类: $\{u_1\}$, $\{u_2, u_3, u_4, u_7\}$, $\{u_5\}$, $\{u_6\}$, $\{u_8\}$ 。
- 当 $\lambda=0.9581$ 时, 分 4 类: $\{u_1\}$, $\{u_2, u_3, u_4, u_6, u_7\}$, $\{u_5\}$, $\{u_8\}$ 。
- 当 $\lambda=0.9545$ 时, 分 3 类: $\{u_1\}$, $\{u_2, u_3, u_4, u_6, u_7, u_8\}$, $\{u_5\}$ 。
- 当 $\lambda=0.7609$ 时, 分 2 类: $\{u_1\}$, $\{u_2, u_3, u_4, u_5, u_6, u_7, u_8\}$ 。
- 当 $\lambda=0.7240$ 时, 分 1 类: $\{u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8\}$ 。

【思考题 5】

设 $X = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$ 是 8 个评价对象, 试给出“严重污染程度”的模糊集表示方法。

2. 设

$$R = \begin{pmatrix} 1.0 & 0.5 & 0.9 \\ 0.5 & 1.0 & 0.5 \\ 0.9 & 0.5 & 1.0 \end{pmatrix}$$

证明 R 是模糊等价矩阵。

3. 设

$$R = \begin{pmatrix} 1.0000 & 0.7000 & 0.9000 & 0.9000 \\ 0.7000 & 1.0000 & 0.7000 & 0.7000 \\ 0.9000 & 0.7000 & 1.0000 & 0.9000 \\ 0.9000 & 0.7000 & 0.9000 & 1.0000 \end{pmatrix}$$

证明 R 是模糊等价矩阵。

4. 试述模糊聚类分析的详细步骤

5. 试用传递闭包法、直接聚类法对矩阵

| | | | | | | | |
|-----|----|-----|-----|-----|-----|-----|-----|
| | | 0.8 | 1.0 | 0.2 | 0.8 | 0.5 | 0.3 |
| | 1. | 0.4 | 0.3 | 0.7 | 0.6 | 0.3 | 1 |
| | 0. | 1.0 | 0.7 | 1.0 | 0.6 | 0.5 | |
| R | 0. | 0.7 | 1.0 | 0.5 | 0.8 | 0.6 | |
| | 0. | 1.0 | 0.5 | 1.0 | 0.2 | 0.7 | |
| | 0. | 0.6 | 0.8 | 0.2 | 1.0 | 0.8 | |
| | 0. | 0.5 | 0.6 | 0.7 | 0.8 | 1.0 | |

作聚类分析, 并作 F 检验

6. 试选一种方法对表 1.7 所示的九个农业区作聚类分析, 并作 F 检验。

地下水水质进行聚类分析, 并作 F 检验

表 1.7 市区及局地下水水质分析样本数据表

| 代号 | 采样地点 | pH 值 | Mg ²⁺ | Cl | SO ₄ ²⁻ | Cl + SO ₄ ²⁻ | HCO ₃ ⁻ | 侵蚀性 CO ₂ |
|----|------|------|------------------|--------|-------------------------------|------------------------------------|-------------------------------|---------------------|
| 1 | 恩华药厂 | 7.30 | 37.16 | 123.56 | 138.06 | 261.62 | 259.25 | 0.00 |
| 2 | 军分区 | 7.50 | 57.39 | 196.37 | 609.99 | 806.36 | 286.06 | 2.64 |
| 3 | 兴隆大厦 | 7.90 | 40.02 | 73.85 | 486.72 | 560.57 | 317.85 | 0.00 |
| 4 | 户部商都 | 6.90 | 34.00 | 115.58 | 252.72 | 368.30 | 496.34 | 0.00 |
| 5 | 九降总部 | 7.60 | 33.17 | 82.07 | 55.36 | 137.43 | 224.37 | 0.00 |
| 6 | 博爱大厦 | 7.50 | 104.38 | 503.63 | 19.14 | 522.77 | 245.12 | 0.00 |
| 7 | 徐州饭店 | 8.15 | 25.06 | 111.34 | 108.05 | 219.39 | 88.09 | 5.99 |
| 8 | 电力宾馆 | 7.65 | 74.32 | 287.78 | 763.01 | 1050.79 | 442.09 | 0.00 |
| 9 | 夹河前街 | 7.50 | 57.34 | 141.46 | 176.58 | 318.04 | 358.97 | 0.00 |
| 10 | 开明市场 | 7.60 | 65.23 | 161.63 | 253.34 | 414.79 | 422.99 | 0.00 |
| 11 | 空后学院 | 8.40 | 16.43 | 83.42 | 226.38 | 309.80 | 442.58 | 0.00 |
| 12 | 灯泡厂 | 7.20 | 66.70 | 151.99 | 219.79 | 371.78 | 784.84 | 0.00 |
| 13 | 府原小区 | 7.70 | 69.25 | 102.53 | 212.40 | 314.93 | 541.80 | 0.00 |
| 14 | 天润花园 | 7.50 | 82.90 | 133.50 | 411.60 | 545.10 | 630.20 | 0.00 |
| 15 | 基督教堂 | 7.60 | 62.54 | 225.98 | 393.69 | 619.67 | 318.71 | 7.14 |
| 16 | 赢都花园 | 7.55 | 11.95 | 91.03 | 293.46 | 384.49 | 287.79 | 2.43 |
| 17 | 交通局 | 7.30 | 35.56 | 41.46 | 64.61 | 106.07 | 443.85 | 0.00 |
| 18 | 望景花园 | 7.30 | 69.78 | 65.86 | 131.06 | 196.92 | 646.57 | 0.00 |
| 19 | 少年巷 | 7.60 | 35.44 | 85.05 | 96.31 | 181.36 | 239.32 | 0.00 |
| 20 | 体育馆 | 7.30 | 47.40 | 129.44 | 142.41 | 271.85 | 191.11 | 0.00 |

第6章 环境判别分析

判别分析是了解客观世界的一种分析方法,而判别分析是在已知分类情况下寻找判别分析的依据。在环境科学中,判别分析常用于环境质量分类、环境影响评价、环境健康危害分类、环境风险评估、环境管理决策等方面。判别分析在环境科学中的应用,对于理解和解决环境问题具有重要意义。判别分析是环境科学中一种重要的工具之一,是解决环境问题的有效工具之一即是判别分析。

判别分析是一种统计方法,用于根据已知分类数据,寻找判别分析的依据。判别分析按照判别分析的数据类型,可以分为线性判别分析和非线性判别分析。线性判别分析适用于线性可分的数据,而非线性判别分析适用于非线性可分的数据。判别分析在环境科学中的应用,对于理解和解决环境问题具有重要意义。判别分析是环境科学中一种重要的工具之一,是解决环境问题的有效工具之一即是判别分析。

本章的主要内容是:

- 距离判别分析;
- Fisher 判别分析;
- Bayes 判别分析;
- 环境应用。

6.1 距离判别分析

距离判别分析,是指根据已知分类数据,寻找判别分析的依据。距离判别分析适用于线性可分的数据,而非线性判别分析适用于非线性可分的数据。距离判别分析在环境科学中的应用,对于理解和解决环境问题具有重要意义。距离判别分析是环境科学中一种重要的工具之一,是解决环境问题的有效工具之一即是距离判别分析。

6.1.1 两总体情况

设有两总体 G_1 和 G_2 , X_1, X_2, \dots, X_n 为 G_1 的样本, Y_1, Y_2, \dots, Y_m 为 G_2 的样本。

的判别函数 $A_1(x)$ 和 $A_2(x)$ 中, $A_1(x) > A_2(x)$ 时, 判别为自污染; 若样本 x 到二体污染函数 $A_1(x)$ 和 $A_2(x)$ 相等, 则判别为自污染和体污染; 反之, 当判别函数 $A_1(x)$ 和 $A_2(x)$ 相等, 且样本 x 到二体污染函数 $A_1(x)$ 和 $A_2(x)$ 相等时, 则判别为自污染和体污染。判别准则的数学模型可描述为:

$$x \in G_1 \quad d(x, G_1) < d(x, G_2)$$

$$A_1(x) > A_2(x) \quad d(x, G_1) < d(x, G_2)$$

$$\text{待判} \quad d(x, G_1) = d(x, G_2)$$

由式(6.1)和式(6.2)可得判别函数 $A_1(x)$ 和 $A_2(x)$ 的表达式, 即

$$d(x, G_1) = (x - \mu_1)' \Sigma_1^{-1} (x - \mu_1) \quad (6.2)$$

$$d(x, G_2) = (x - \mu_2)' \Sigma_2^{-1} (x - \mu_2) \quad (6.3)$$

式中, $\mu = \mu_1, \mu_2$; $\Sigma = \Sigma_1, \Sigma_2$ 分别为二体污染函数 $A_1(x)$ 和 $A_2(x)$ 的协方差矩阵 Σ_1 和 Σ_2 的逆矩阵。

由式(6.2)和式(6.3)可得判别函数 $A_1(x)$ 和 $A_2(x)$ 的表达式, 即

$$A_1(x) = (x - \mu_1)' \Sigma_1^{-1} (x - \mu_1)$$

$$A_2(x) = (x - \mu_2)' \Sigma_2^{-1} (x - \mu_2)$$

判别函数 $A_1(x)$ 和 $A_2(x)$ 的表达式, 即

由式(6.2)和式(6.3)可得判别函数 $A_1(x)$ 和 $A_2(x)$ 的表达式, 即

$$d(x, G_1) - d(x, G_2) = -2[x - (\mu_1 + \mu_2)/2]' \Sigma^{-1} (\mu_1 - \mu_2)$$

令

$$\bar{\mu} = (\mu_1 + \mu_2)/2$$

则

$$W(x) = (x - \bar{\mu})' \Sigma^{-1} (\mu_1 - \mu_2) \quad (6.4)$$

于是判别规则(6.1)可以表示为:

$$\begin{cases} x \in G_1 & W(x) > 0 \\ x \in G_2 & W(x) < 0 \\ \text{待判} & W(x) = 0 \end{cases} \quad (6.5)$$

式中, $W(x)$ 为判别函数, $\mu_1 = \bar{x}_1$ 为自污染函数 $A_1(x)$ 的均值, $\mu_2 = \bar{x}_2$ 为体污染函数 $A_2(x)$ 的均值, Σ 为判别函数 $A_1(x)$ 和 $A_2(x)$ 的协方差矩阵 Σ_1 和 Σ_2 的逆矩阵。

由式(6.4)可得判别函数 $A_1(x)$ 和 $A_2(x)$ 的表达式, 即

$$A_1(x) = (x - \mu_1)' \Sigma_1^{-1} (x - \mu_1)$$

$$A_2(x) = (x - \mu_2)' \Sigma_2^{-1} (x - \mu_2)$$

判别函数 $A_1(x)$ 和 $A_2(x)$ 的表达式, 即

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$$

$$\hat{\mu}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} y_i = \bar{y}$$

$$\Sigma = \frac{1}{n_1 + n_2} (A_1 + A_2)$$

$$\begin{aligned} \text{其中, } A_1 &= \sum_{j=1}^n (x_j - \bar{x})(x_j - \bar{x})'; A_2 = \sum_{j=1}^n (y_j - \bar{y})(y_j - \bar{y})' \\ &= (x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x})' \Sigma (x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x}) \\ &= d(x, G_1) - d(x, G_2) \\ &= (x - \mu_1)' \Sigma_1^{-1} (x - \mu_1) - (x - \mu_2)' \Sigma_2^{-1} (x - \mu_2) \\ &= (x - \mu)' \Sigma^{-1} (x - \mu) \quad (\mu = \frac{1}{2}(\mu_1 + \mu_2), \Sigma = \frac{1}{2}(\Sigma_1 + \Sigma_2)) \end{aligned}$$

$$\bar{\Sigma}_m = \frac{1}{n_m - 1} A_m \quad (m=1, 2)$$

式中, $A_m = \sum_{i=1}^n (x_i^{(m)} - \bar{x}^{(m)})(x_i^{(m)} - \bar{x}^{(m)})'$ 为第 m 类样本的离差阵, $\bar{x}^{(m)}$ 为第 m 类样本的均值向量, Σ_m 为第 m 类样本的协方差阵, Σ 为两类的协方差阵, μ 为两类的均值向量, μ_1 和 μ_2 分别为两类的均值向量, Σ_1 和 Σ_2 分别为两类的协方差阵。

(1) 这种判别是符合习惯的。

如图 6-1 所示, 设 μ_1 和 μ_2 为两类的均值向量, Σ_1 和 Σ_2 分别为两类的协方差阵, Σ 为两类的协方差阵, μ 为两类的均值向量。当 Σ_1 和 Σ_2 相等时, Σ 为 Σ_1 和 Σ_2 的平均值, 此时判别规则为: 若 $d(x, G_1) < d(x, G_2)$, 则判 x 属于 G_1 ; 若 $d(x, G_1) > d(x, G_2)$, 则判 x 属于 G_2 ; 若 $d(x, G_1) = d(x, G_2)$, 则判 x 属于 G_1 和 G_2 的交集。

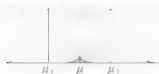


图 6-1

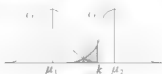


图 6-2

如图 6-2 所示, 设 μ_1 和 μ_2 为两类的均值向量, Σ_1 和 Σ_2 分别为两类的协方差阵, Σ 为两类的协方差阵, μ 为两类的均值向量。当 Σ_1 和 Σ_2 不相等时, Σ 为 Σ_1 和 Σ_2 的平均值, 此时判别规则为: 若 $d(x, G_1) < d(x, G_2)$, 则判 x 属于 G_1 ; 若 $d(x, G_1) > d(x, G_2)$, 则判 x 属于 G_2 ; 若 $d(x, G_1) = d(x, G_2)$, 则判 x 属于 G_1 和 G_2 的交集。

如图 6-3 所示, 设 μ_1 和 μ_2 为两类的均值向量, Σ_1 和 Σ_2 分别为两类的协方差阵, Σ 为两类的协方差阵, μ 为两类的均值向量。当 Σ_1 和 Σ_2 不相等时, Σ 为 Σ_1 和 Σ_2 的平均值, 此时判别规则为: 若 $d(x, G_1) < d(x, G_2)$, 则判 x 属于 G_1 ; 若 $d(x, G_1) > d(x, G_2)$, 则判 x 属于 G_2 ; 若 $d(x, G_1) = d(x, G_2)$, 则判 x 属于 G_1 和 G_2 的交集。

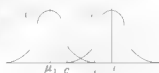


图 6-3

$$\begin{aligned}x &\in G_1 & (x \leq c) \\x &\in G_2 & (x \geq d) \\&\text{待判} & (c < x < d)\end{aligned}$$

综上所述,距离判别分析的步骤如下:

- (1)估计总体 G_1 和 G_2 的均值、协方差矩阵;
- (2)计算判别函数 $W(x)$;
- (3)把待判样本代入 $W(x)$ 进行判断。

$$\begin{aligned}x &\in G_1 & (W(x) > 0) \\x &\in G_2 & (W(x) < 0) \\&\text{待判} & (W(x) = 0)\end{aligned}$$

例 6.1 某地区有 10 种大气污染物, 其中 5 种为有害气体, 5 种为颗粒物。现对 10 种大气污染物进行判别分析, 判别函数 $W(x)$ 如下: $W(x) = 0.1x_1 + 0.2x_2 + 0.3x_3 + 0.4x_4 + 0.5x_5 + 0.6x_6 + 0.7x_7 + 0.8x_8 + 0.9x_9 + 1.0x_{10}$, 其中 x_1, x_2, x_3, x_4, x_5 为有害气体, $x_6, x_7, x_8, x_9, x_{10}$ 为颗粒物。现有一待判样本, 其判别函数值为 $W(x) = 0.5$, 问该样本属于哪类。

表 6.1 两种大气污染物下的植物反应

| 组别 | 序号 | 叶色指数 x_1 | 植株生长指数 x_2 |
|-------------------------------|----|------------|--------------|
| 第一组
遭受 SO_2
污染 | 1 | 9.6 | 19.6 |
| | 2 | 9.3 | 19.9 |
| | 3 | 8.7 | 18.6 |
| | 4 | 8.8 | 18.9 |
| | 5 | 8.5 | 19.6 |
| | 6 | 10.2 | 30.3 |
| 第二组
遭受 HCl
污染 | 1 | 11.3 | 28.7 |
| | 2 | 9.8 | 25.6 |
| | 3 | 7.2 | 27.6 |
| | 4 | 8.5 | 29.0 |
| | 5 | 9.6 | 30.6 |
| | 6 | 9.2 | 19.0 |
| 待判样本 | 1 | 9.2 | 19.0 |
| | 2 | 8.6 | 19.6 |
| | 3 | 11.2 | 30.3 |

解 由表 6.1 可知, 第一组植物反应为 x_1, x_2 , 第二组植物反应为 x_3, x_4 , 待判样本植物反应为 x_5, x_6 。计算各组的平均指标和协方差

$$\bar{x}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i = \bar{x} = (8.980 \quad 19.320)$$

$$\bar{\mu}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} y_i = \mathbf{y} = (9.433 \ 3 \quad 28.533 \ 3)$$

$$\boldsymbol{\mu} = (9.206 \ 7 \quad 23.926 \ 7)$$

总体协方差矩阵和它的逆矩阵为:

$$\hat{\Sigma} = \begin{bmatrix} 1.213 \ 5 & 0.331 \ 7 \\ 0.331 \ 7 & 1.797 \ 9 \end{bmatrix}$$

$$\hat{\Sigma}^{-1} = \begin{bmatrix} 0.867 \ 8 & -0.160 \ 1 \\ -0.160 \ 1 & 0.585 \ 7 \end{bmatrix}$$

$$\hat{\Sigma}^{-1}(\bar{\mu}_1 - \bar{\mu}_2) = (1.081 \ 7 \quad -5.324 \ 0)'$$

从而判别函数:

$$\begin{aligned} W(x) &= (x - \boldsymbol{\mu})' \hat{\Sigma}^{-1}(\bar{\mu}_1 - \bar{\mu}_2) \\ &= 1.081 \ 7 (x_1 - 9.206 \ 7) - 5.324 \ 0 (x_2 - 23.926 \ 7) \end{aligned}$$

将三个样本代入判别函数, 可得: $W_1 = 26.222 \ 3$, $W_2 = -22.378 \ 9$, $W_3 = -31.775 \ 3$

$$W_1 = 26.222 \ 3, W_2 = -22.378 \ 9, W_3 = -31.775 \ 3$$

因为 $W_1 > 0$, $W_2 < 0$, $W_3 < 0$, 所以样本 1 属于 G_1 , 样本 2 属于 G_2 , 样本 3 属于 G_2 。

6.1.2 多总体情况

对于多总体判别问题, 在判别多总体情况时, 我们一般从判别函数在相同和不相同两个方面来考虑。

6.1.2.1 协方差矩阵相同

设有 k 个总体 G_1, G_2, \dots, G_k , 它们的均值分别为 $\mu_1, \mu_2, \dots, \mu_k$, 协方差矩阵为 Σ , 则判别函数为:

$$W_{ij}(x) = \left(x - \frac{\mu_i + \mu_j}{2} \right)' \Sigma^{-1}(\mu_i - \mu_j) \quad (i, j=1, 2, \dots, k)$$

相应的判别规则为:

$$\begin{cases} x \in G_i & (W_{ij}(x) > 0, \forall j \neq i) \\ \text{待判} & (\text{某个 } W_{ij}(x) = 0) \end{cases}$$

当 $\mu_1, \mu_2, \dots, \mu_k \in \Sigma^{-1/2} \cdot \mathbb{R}^k$ 时, 设 $x \in \mathbb{R}^k$ 且取 x 和 μ_i 的距离为 $\lambda_1, \lambda_2, \dots, \lambda_k$, $x \in G_m$ ($m=1, 2, \dots, k$), 则它们的估计为:

$$\begin{aligned}
 & \mathbf{X}^{(1)} = \begin{pmatrix} x_{11}^{(1)} & x_{12}^{(1)} & \cdots & x_{1n_1}^{(1)} \\ x_{21}^{(1)} & x_{22}^{(1)} & \cdots & x_{2n_1}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1}^{(1)} & x_{p2}^{(1)} & \cdots & x_{pn_1}^{(1)} \end{pmatrix} \\
 & \mathbf{X}^{(2)} = \begin{pmatrix} x_{11}^{(2)} & x_{12}^{(2)} & \cdots & x_{1n_2}^{(2)} \\ x_{21}^{(2)} & x_{22}^{(2)} & \cdots & x_{2n_2}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1}^{(2)} & x_{p2}^{(2)} & \cdots & x_{pn_2}^{(2)} \end{pmatrix}
 \end{aligned}$$

令 $\mathbf{X} = (x_{11}, x_{12}, \dots, x_{1n_1}, x_{21}, x_{22}, \dots, x_{2n_1}, \dots, x_{p1}, x_{p2}, \dots, x_{pn_1}, x_{11}, x_{12}, \dots, x_{1n_2}, x_{21}, x_{22}, \dots, x_{2n_2}, \dots, x_{p1}, x_{p2}, \dots, x_{pn_2})$ 为 $n_1 + n_2$ 个变量组成的行向量, $\mathbf{X}^T = (x_{11}, x_{21}, \dots, x_{p1}, x_{12}, x_{22}, \dots, x_{pn_2})$ 为 $n_1 + n_2$ 个变量组成的列向量, 则 $\mathbf{X} \mathbf{X}^T$ 为 $(n_1 + n_2) \times (n_1 + n_2)$ 阶方阵, 其主对角线元素为 $x_{11}^2, x_{21}^2, \dots, x_{p1}^2, x_{12}^2, x_{22}^2, \dots, x_{pn_2}^2$, 其余元素为 $x_{ij}x_{kl}$ 。

$$y_j^{(j)} = c_1 x_{1j}^{(j)} + c_2 x_{2j}^{(j)} + \cdots + c_p x_{pj}^{(j)} \quad (j=1, 2, \dots, n_2)$$

令:

$$\mathbf{C} = (c_1, c_2, \dots, c_p)$$

对于分类来说, 显然要求:

$$Q = \sum_{j=1}^{n_1} (x_{1j} - c_1)^2 + \sum_{j=1}^{n_2} (x_{1j} - c_1)^2 + \sum_{j=1}^{n_2} (x_{2j} - c_2)^2 + \cdots + \sum_{j=1}^{n_2} (x_{pj} - c_p)^2$$

$$Q = \sum_{j=1}^{n_1} (x_{1j} - c_1)^2 + \sum_{j=1}^{n_2} (x_{1j} - c_1)^2 + \sum_{j=1}^{n_2} (x_{2j} - c_2)^2 + \cdots + \sum_{j=1}^{n_2} (x_{pj} - c_p)^2$$

好。记:

$$F = \sum_{j=1}^{n_1} (y_j^{(1)} - \bar{y}^{(1)})^2 + \sum_{j=1}^{n_2} (y_j^{(2)} - \bar{y}^{(2)})^2$$

根据上述两点准则, 我们应该使

$$F = \frac{Q}{J}$$

越大越好。

对上式两边分取对数有:

$$\ln F = \ln Q - \ln J$$

令 $F = \ln Q - \ln J$, 则 F 为 c_1, c_2, \dots, c_p 的函数, 令 F 对 c_1, c_2, \dots, c_p 满足极值条件, 应有

$$I = \frac{1}{n} \sum_{j=1}^n Q_j = \frac{1}{n} I$$

即:

$$Q = \frac{1}{n} I$$

所以:

$$Q = \frac{1}{n} (Q_1 + Q_2 + \dots + Q_p)$$

;

$$\begin{aligned} &= \sum_{j=1}^n c_1 x_j^{(1)} + \sum_{j=1}^n c_2 x_j^{(2)} + \dots + \sum_{j=1}^n c_p x_j^{(p)} \\ &= c_1 \bar{x}_1^{(1)} + c_2 \bar{x}_2^{(1)} + \dots + c_p \bar{x}_p^{(1)} \end{aligned}$$

从而:

$$\begin{aligned} F &= \sum_{j=1}^{n_1} (y_j^{(1)} - \bar{y}^{(1)})^2 + \sum_{j=1}^{n_2} (y_j^{(2)} - \bar{y}^{(2)})^2 \\ &= \sum_{j=1}^{n_1} \sum_{k=1}^p c_k (x_{kj}^{(1)} - \bar{x}_k^{(1)})^2 + \sum_{j=1}^{n_2} \sum_{k=1}^p c_k (x_{kj}^{(2)} - \bar{x}_k^{(2)})^2 \\ &= \sum_{j=1}^{n_1} \sum_{k=1}^p c_k (x_{kj}^{(1)} - \bar{x}_k^{(1)})^2 + \sum_{j=1}^{n_2} \sum_{k=1}^p c_k (x_{kj}^{(2)} - \bar{x}_k^{(2)})^2 \\ &= \sum_{k=1}^p c_k \sum_{j=1}^{n_1} (x_{kj}^{(1)} - \bar{x}_k^{(1)})^2 + \sum_{k=1}^p c_k \sum_{j=1}^{n_2} (x_{kj}^{(2)} - \bar{x}_k^{(2)})^2 \\ &= \sum_{k=1}^p c_k \sum_{j=1}^{n_1} (x_{kj}^{(1)} - \bar{x}_k^{(1)}) (x_{kj}^{(1)} - \bar{x}_k^{(1)}) + \\ &\quad \sum_{k=1}^p c_k \sum_{j=1}^{n_2} (x_{kj}^{(2)} - \bar{x}_k^{(2)}) (x_{kj}^{(2)} - \bar{x}_k^{(2)}) \\ &= \sum_{k=1}^p \sum_{j=1}^{n_1} c_k c_{kj} \end{aligned}$$

这里:

$$l_{be} = \sum_{j=1}^{n_1} (x_{bj}^{(1)} - \overline{x_b^{(1)}})(x_{ej}^{(1)} - \overline{x_e^{(1)}}) + \sum_{j=1}^{n_2} (x_{bj}^{(2)} - \overline{x_b^{(2)}})(x_{ej}^{(2)} - \overline{x_e^{(2)}})$$

$$l_{ke} - l_{eb}$$

并且有:

$$\sum_{j=1}^{n_1} (x_{bj}^{(1)} - \overline{x_b^{(1)}})^2 = \sum_{j=1}^{n_2} (x_{bj}^{(2)} - \overline{x_b^{(2)}})^2 = \sum_{j=1}^{n_1} (x_{bj}^{(1)} - \overline{x_b^{(1)}})(x_{bj}^{(2)} - \overline{x_b^{(2)}})$$

当 $k = e$ 时, 上式 = $\sum_{j=1}^{n_1} (x_{bj}^{(1)})^2 - \frac{1}{n_1} \sum_{j=1}^{n_1} x_{bj}^{(1)}$

对 $x^{(2)}$ 也有同样结果。

所以有:

$$\frac{\partial F}{\partial c_k} = 2(c_1 l_{k1} + c_2 l_{k2} + \cdots + c_p l_{kp}) \quad (k=1, 2, \cdots, p)$$

$$Q = \sum_{r=1}^p c_r (\overline{x_r^{(1)}} - \overline{x_r^{(2)}}) \quad \dots$$

$$= \left[\sum_{r=1}^p c_r (\overline{x_r^{(1)}} - \overline{x_r^{(2)}}) \right] = \left(\sum_{r=1}^p c_r t_r \right)$$

其中

$$t_r = \overline{x_r^{(1)}} - \overline{x_r^{(2)}} \quad (r=1, 2, \cdots, p)$$

由此有:

$$\frac{\partial Q}{\partial c_k} = 2 \left(\sum_{r=1}^p c_r t_r \right) t_k \quad (k=1, 2, \cdots, p)$$

$$I \frac{\partial F}{\partial c_k} = \frac{\partial Q}{\partial c_k}$$

$$I(c_1 l_{k1} + c_2 l_{k2} + \cdots + c_p l_{kp}) = \left(\sum_{r=1}^p c_r t_r \right) t_k \quad (k=1, 2, \cdots, p)$$

$$\sum_{r=1}^p$$

则有:

$$\cdots \quad \dots$$

因而有:

$$c_1 l_{11} + c_2 l_{12} + \cdots + c_p l_{1p} = \beta_1$$

$$\vdots$$

$$c_1 l_{p1} + c_2 l_{p2} + \cdots + c_p l_{pp} = \beta_p$$

上面方程组的解是如下方程组解的 β 倍:

$$\sum_{j=1}^p c_j l_{kj} = t_k \quad (k=1, 2, \cdots, p)$$

$$I = \frac{Q}{I} = (Q, t_1, \cdots, t_p) \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = Q$$

$$= c_1 t_1 + c_2 t_2 + \cdots + c_p t_p = \sum_{j=1}^p c_j t_j = \sum_{j=1}^p c_j \sum_{k=1}^p c_k l_{kj} = \sum_{k=1}^p \sum_{j=1}^p c_k c_j l_{kj} = \sum_{k=1}^p \sum_{j=1}^p c_k c_j l_{jk}$$

可使 I 达到最大值。这样, 作为线性函数:

$$y = c_1 x_1 + c_2 x_2 + \cdots + c_p x_p$$

就是我们所要求的, 被称为判别函数

$$y = c_1 x_1 + c_2 x_2 + \cdots + c_p x_p = \sum_{j=1}^p c_j x_j = \sum_{j=1}^p c_j \sum_{k=1}^p c_k l_{kj} = \sum_{k=1}^p \sum_{j=1}^p c_k c_j l_{kj} = \sum_{k=1}^p \sum_{j=1}^p c_k c_j l_{jk}$$

$+ n_2 - 2$ 。因而 y 的选择有:

$$y_0 = \frac{1}{2}(\hat{\mu}_1 - \hat{\mu}_2)' \hat{\Sigma}^{-1}(\hat{\mu}_1 + \hat{\mu}_2)$$

$$y_0 = \frac{1}{2}(\hat{\mu}_1 - \hat{\mu}_2)' \hat{\Sigma}^{-1}(\hat{\mu}_1 + \hat{\mu}_2)$$

不失一般性, 假设:

若 $(x_1 - \hat{\mu}_1)' \hat{\Sigma}^{-1}(x_1 - \hat{\mu}_1) < (x_1 - \hat{\mu}_2)' \hat{\Sigma}^{-1}(x_1 - \hat{\mu}_2)$, 则判为第一类, 否则属于第二类。

Fisher 判别分析的步骤如下:

(1) 估计总体 G_1, G_2 的均值和协方差矩阵 $\hat{\mu}_1, \hat{\mu}_2, \Sigma$;

(2) 计算 $y_0 = \frac{1}{2}(\hat{\mu}_1 - \hat{\mu}_2)' \hat{\Sigma}^{-1}(\hat{\mu}_1 + \hat{\mu}_2)$;

(3) 计算 $c = \Sigma^{-1}(\hat{\mu}_1 - \hat{\mu}_2)$;

(4) 求判别函数 $y = x'c$;

$$y = c_1 x_1 + c_2 x_2 + \cdots + c_p x_p = \sum_{j=1}^p c_j x_j = \sum_{j=1}^p c_j \sum_{k=1}^p c_k l_{kj} = \sum_{k=1}^p \sum_{j=1}^p c_k c_j l_{kj} = \sum_{k=1}^p \sum_{j=1}^p c_k c_j l_{jk}$$

(6) 待判样本的判别。

判别函数为 $y_0 = -117.4263 - 1.0817x_1 - 5.3240x_2$ 。将待判样本代入判别函数, 确定待判样本的归属。

例 6.2 某地区有 3 种类型的土壤, 其分析数据列于表 6.2。

解 根据距离判别中的分析数据, 可以得到:

$$y_0 = \frac{1}{2} (\hat{\mu}_1 - \hat{\mu}_2)' \hat{\Sigma}^{-1} (\hat{\mu}_1 + \hat{\mu}_2) = -117.4263$$

$$y - c'x = (\hat{\mu}_1 - \hat{\mu}_2)' \hat{\Sigma}^{-1} x = 1.0817x_1 + 5.3240x_2$$

$$\overline{y^{(1)}} = \frac{1}{n_1} \sum_{j=1}^n y_j^{(1)} = -93.1457$$

$$y^{(2)} = \frac{1}{n_2} \sum_{j=1}^n y_j^{(2)} = -141.7070$$

$$\text{即 } \overline{y^{(1)}} > y_0 > \overline{y^{(2)}}$$

将样本 1、2、3 的数据分别代入到判别函数中, 得到:

$$y_1 = 91.2010, y_2 = -95.0474, y_3 = 149.2017$$

因此, 样本 1 属于第一类土壤, 样本 2 属于第二类土壤, 样本 3 属于第三类土壤。这个结果和距离判别的结果是一致的。

6.3 Bayes 判别

Bayes 判别法是根据 Bayes 公式, 利用先验概率和条件概率, 计算后验概率, 从而对样本进行分类。Bayes 判别法的基本思想是: 假设样本属于第 i 类, 其先验概率为 $P(G_i)$, 条件概率为 $P(x|G_i)$, 则后验概率为 $P(G_i|x)$ 。Bayes 判别法的基本公式为:

假设样本 x 属于第 i 类, 其先验概率为 $P(G_i)$, 条件概率为 $P(x|G_i)$, 则后验概率为 $P(G_i|x)$ 。Bayes 判别法的基本公式为:

假设样本 x 属于第 i 类, 其先验概率为 $P(G_i)$, 条件概率为 $P(x|G_i)$, 则后验概率为 $P(G_i|x)$ 。Bayes 判别法的基本公式为:

$$P(G_i | x) = \frac{P(G_i)P(x | G_i)}{\sum_{i=1}^2 P(x | G_i)P(G_i)} = \frac{q_i P(x | G_i)}{q_1 P(x | G_1) + q_2 P(x | G_2)} \quad (i=1, 2) \quad (6.6)$$

由式(6.6)可知, 判别函数 $P(G_i | x)$ 与 $P(G_i)$ 成正比, 因此, 在判别函数 $P(G_i | x)$ 中, 只要知道 $P(x | G_i)$ 即可。因此, 在判别函数 $P(G_i | x)$ 中, 只要知道 $P(x | G_i)$ 即可。

$$P(G_i | x) = q_i p_i(x) + q_j p_j(x) \quad (i=1, 2) \quad (6.7)$$

把上式代入 $P(G_1 | x) \geq P(G_2 | x)$, $P(G_1 | x) < P(G_2 | x)$ 就有:

若 $\frac{q_1 p_1(x)}{q_2 p_2(x)} \geq 1$ 事件 x 划归 G_1 总体;

若 $\frac{q_1 p_1(x)}{q_2 p_2(x)} < 1$ 事件 x 划归 G_2 总体。

因此, 判别函数 $P(G_i | x)$ 与 $P(G_i)$ 成正比, 因此, 在判别函数 $P(G_i | x)$ 中, 只要知道 $P(x | G_i)$ 即可。

$$q_1 P(2 | 1) + q_2 P(1 | 2) \quad (6.8)$$

由式(6.8)可知, 判别函数 $P(G_i | x)$ 与 $P(G_i)$ 成正比, 因此, 在判别函数 $P(G_i | x)$ 中, 只要知道 $P(x | G_i)$ 即可。

因此, 判别函数 $P(G_i | x)$ 与 $P(G_i)$ 成正比, 因此, 在判别函数 $P(G_i | x)$ 中, 只要知道 $P(x | G_i)$ 即可。

$$q_1 L(2 | 1) P(2 | 1) + q_2 L(1 | 2) P(1 | 2) \quad (6.9)$$

因此, 判别函数 $P(G_i | x)$ 与 $P(G_i)$ 成正比, 因此, 在判别函数 $P(G_i | x)$ 中, 只要知道 $P(x | G_i)$ 即可。

因此, 判别函数 $P(G_i | x)$ 与 $P(G_i)$ 成正比, 因此, 在判别函数 $P(G_i | x)$ 中, 只要知道 $P(x | G_i)$ 即可。

因此, 判别函数 $P(G_i | x)$ 与 $P(G_i)$ 成正比, 因此, 在判别函数 $P(G_i | x)$ 中, 只要知道 $P(x | G_i)$ 即可。

$$P(j | i) = \int p_i(x) dx \quad (i \neq j, j=1, 2, \dots, m)$$

因此, 判别函数 $P(G_i | x)$ 与 $P(G_i)$ 成正比, 因此, 在判别函数 $P(G_i | x)$ 中, 只要知道 $P(x | G_i)$ 即可。

$$g(D_1, D_2, \dots, D_m) = \sum_{i=1}^m q_i \sum_{j=1}^m L(j | i) P(j | i) \quad (6.10)$$

可以证明,当某种划分 D_1, D_2, \dots, D_m 满足:

$$D_l = \{x: h_l(x) < h_j(x) \quad (j \neq l; j=1, 2, \dots, m; l=1, 2, \dots, m)\}$$

时,能使平均损失 $g(D_1, D_2, \dots, D_m)$ 取得最小值。其中:

$$h_j(x) = \sum_{i=1}^m q_i L(j-i) p_i(x) \quad (6.11)$$

由式(6.11)可以看出,要计算 $h_j(x)$, 必须先知道 $p_i(x)$ 的表达式,而 $p_i(x)$ 的表达式又不易求出,所以往往假设 $L(j-i)=1$ 。这样:

$$h_j(x) = \sum_{i=1}^m q_i p_i(x) = \sum_{i=1}^m q_i \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_i)' \Sigma^{-1}(x-\mu_i)\right]$$

于是, $h_j(x)$ 的表达式就变为: $h_j(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_j)' \Sigma^{-1}(x-\mu_j)\right]$ 。同时, $p_i(x)$ 的表达式也变为: $p_i(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_i)' \Sigma^{-1}(x-\mu_i)\right]$ 。因此,式(6.11)变为:

$$p_i(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_i)' \Sigma^{-1}(x-\mu_i)\right] \quad (6.13)$$

于是, $h_j(x)$ 的表达式变为: $h_j(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_j)' \Sigma^{-1}(x-\mu_j)\right]$ 。因此,式(6.11)变为:

$$y_i(x) = c_0 + c_{10}x_1 + c_{20}x_2 + \dots + c_{p0}x_p + \ln q_i \quad (i=1, 2, \dots, m) \quad (6.14)$$

于是, $y_i(x)$ 的表达式变为: $y_i(x) = c_0 + c_{10}x_1 + c_{20}x_2 + \dots + c_{p0}x_p + \ln q_i$ 。因此,式(6.14)变为:

$$D_i = \{x | y_i(x) = \max_{j=1,2,\dots,m} y_j(x)\}$$

于是, $y_i(x)$ 的表达式变为: $y_i(x) = c_0 + c_{10}x_1 + c_{20}x_2 + \dots + c_{p0}x_p + \ln q_i$ 。因此,式(6.14)变为:

于是, $y_i(x)$ 的表达式变为: $y_i(x) = c_0 + c_{10}x_1 + c_{20}x_2 + \dots + c_{p0}x_p + \ln q_i$ 。因此,式(6.14)变为:

$$y_i(x) = c_0 + c_{10}x_1 + c_{20}x_2 + \dots + c_{p0}x_p \quad (i=1, 2, \dots, m)$$

具体在计算系数时,可以证明:

$$y_i = \ln q_i + \mu_i' \Sigma^{-1} x - \frac{1}{2} \mu_i' \Sigma^{-1} \mu_i \quad (i=1, 2, \dots, m) \quad (6.15)$$

于是, $y_i(x)$ 的表达式变为: $y_i(x) = c_0 + c_{10}x_1 + c_{20}x_2 + \dots + c_{p0}x_p + \ln q_i$ 。因此,式(6.14)变为:

为第 i 类的样本数。

Bayes 判别分析的步骤如下:

(1) 估计总体的均值和协方差阵:

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}, \quad \Sigma_i = \frac{1}{n_i} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)^T, \quad i = 1, 2, \dots, g;$$

(3) 待判样本的判别

对于待判样本 x , 计算判别函数值 $W_i(x)$, 并比较 $W_i(x)$ 的大小, 将 x 判为 G_i 类。

6.4 环境应用

例 6.3 贝叶斯判别分析理论在安全评价中的应用

某厂生产的产品, 因受环境因素的影响, 产品质量不稳定, 现拟用贝叶斯判别分析理论, 对产品质量进行判别。已知该厂生产的产品, 其质量指标 X_1, X_2, X_3, X_4, X_5 服从多元正态分布, 且各指标均服从正态分布, 其均值和协方差阵如下表所示。现有一批产品, 其质量指标 X_1, X_2, X_3, X_4, X_5 分别为 95, 4, 18, 2.5, 1.2, 试判别该批产品的质量等级。

表 6.2

安全评价等级表

| 指标 | 安全评价等级 | | | | |
|--|---------|---------|---------|---------|--------|
| | 安全(5) | 较安全(4) | 一般安全(3) | 较不安全(2) | 不安全(1) |
| 巷道合格率 % | >95 | 90~95 | 85~90 | 80~85 | <80 |
| 粉尘浓度 ($\text{mg} \cdot \text{m}^{-3}$) | <4 | 4~6 | 6~8 | 8~10 | >10 |
| 环境温度/ $^{\circ}\text{C}$ | 18~22 | 22~24 | 24~26 | 26~28 | >28 |
| 风速 ($\text{m} \cdot \text{s}^{-1}$) | 2.5~3.5 | 2.0~2.5 | 1.5~2.0 | 1.5~1.0 | <1.0 |
| 巷道最小行人宽度/m | >1.2 | 1.1~1.2 | 1.0~1.1 | 0.8~1.0 | <0.8 |
| 巷道最小行人高度/m | >1.8 | 1.6~1.8 | 1.4~1.6 | 1.2~1.4 | <1.2 |

表 6.3 原始数据

| 矿井编号 | X_1 | X_2 | X_3 | X_4 | X_5 | X_6 | 期望值 |
|------|-------|-------|-------|-------|-------|-------|-----|
| 1 | 97.38 | 2.12 | 21.5 | 2.87 | 1.40 | 1.83 | 5 |
| 2 | 98.10 | 3.65 | 19.5 | 3.35 | 1.31 | 2.24 | 5 |
| 3 | 96.45 | 3.14 | 18.0 | 3.50 | 1.20 | 1.94 | 5 |
| 4 | 95.30 | 3.87 | 22.0 | 2.56 | 1.25 | 2.50 | 5 |
| 5 | 94.87 | 4.03 | 23.1 | 2.01 | 1.17 | 1.79 | 4 |
| 6 | 93.15 | 5.35 | 22.7 | 2.32 | 1.19 | 1.72 | 4 |
| 7 | 91.57 | 4.89 | 22.2 | 2.21 | 1.13 | 1.68 | 4 |
| 8 | 90.78 | 5.87 | 23.8 | 2.48 | 1.10 | 1.60 | 4 |
| 9 | 87.69 | 6.17 | 25.9 | 1.98 | 1.05 | 1.47 | 3 |
| 10 | 89.34 | 7.32 | 24.3 | 1.55 | 1.07 | 1.52 | 3 |
| 11 | 85.10 | 6.87 | 25.2 | 1.63 | 1.09 | 1.59 | 3 |
| 12 | 86.54 | 7.91 | 24.0 | 1.75 | 1.00 | 1.41 | 3 |
| 13 | 84.68 | 8.07 | 26.1 | 1.50 | 0.86 | 1.39 | 2 |
| 14 | 84.10 | 9.13 | 27.9 | 1.48 | 0.91 | 1.28 | 2 |
| 15 | 82.34 | 8.63 | 27.4 | 1.35 | 0.97 | 1.35 | 2 |
| 16 | 80.25 | 9.87 | 26.5 | 1.14 | 0.81 | 1.20 | 2 |
| 17 | 75.68 | 10.05 | 30.7 | 0.56 | 0.79 | 1.89 | 1 |
| 18 | 78.98 | 12.30 | 28.9 | 0.87 | 0.45 | 0.78 | 1 |
| 19 | 73.56 | 11.28 | 29.6 | 0.96 | 0.58 | 0.98 | 1 |
| 20 | 70.14 | 10.87 | 28.5 | 0.48 | 0.74 | 1.17 | 1 |

解 根据表中的数据, 得到:

$$\hat{\mu}_1 (96.8075 \quad 3.1950 \quad 20.2500 \quad 3.0700 \quad 1.2900 \quad 2.1275)$$

$$\hat{\mu}_2 (92.5925 \quad 5.0350 \quad 22.9500 \quad 2.2550 \quad 1.1475 \quad 1.6975)$$

$$\hat{\mu}_3 (87.1675 \quad 7.0675 \quad 24.8500 \quad 1.7275 \quad 1.0525 \quad 1.4975)$$

$$\hat{\mu}_4 (82.8425 \quad 8.9250 \quad 26.9750 \quad 1.3675 \quad 0.8875 \quad 1.3050)$$

$$\mu_5 (74.5900 \quad 11.1250 \quad 29.4250 \quad 0.7175 \quad 0.6400 \quad 1.2050)$$

$$\hat{\Sigma} \begin{pmatrix} 5.1129 & -0.2132 & 0.0066 & 0.1467 & 0.0376 & -0.0374 \\ -0.2132 & 0.6430 & 0.2012 & 0.0233 & -0.0474 & 0.0697 \\ 0.0066 & 0.2012 & 1.2457 & 0.1179 & 0.0381 & 0.1122 \\ 0.1467 & 0.0233 & 0.1179 & 0.0687 & -0.0088 & 0.0288 \\ 0.0376 & 0.0474 & 0.0381 & -0.0088 & 0.0079 & 0.0127 \\ 0.0374 & 0.0697 & 0.1122 & -0.0288 & 0.0127 & 0.0689 \end{pmatrix}$$

$$\hat{\Sigma} = \begin{bmatrix} 0.2379 & 0.2675 & -0.0700 & 0.4287 & 2.9515 & -0.2091 \\ 0.2675 & 3.2278 & 0.0498 & 1.0348 & 22.5338 & 0.2288 \\ -0.0700 & -0.0498 & 1.0824 & 1.3111 & 3.2848 & 0.6976 \\ 0.4287 & 1.0348 & 1.3111 & 21.1819 & 13.3569 & 5.0551 \\ 2.9515 & 22.5338 & -3.2848 & 13.3569 & 357.7665 & 30.5919 \\ -0.2091 & -0.2288 & 0.6976 & 5.0551 & -30.5919 & 23.0396 \end{bmatrix}$$

$$\hat{\mu}'\hat{\Sigma}^{-1} = (24.5171 \quad 66.9584 \quad 13.2904 \quad 81.3633 \quad 728.6500 \quad 10.0293)'$$

$$\hat{\mu}_1'\hat{\Sigma}^{-1} = (23.8363 \quad 67.6796 \quad 16.1156 \quad 67.2740 \quad 700.0895 \quad -21.1198)'$$

$$\hat{\mu}_2'\hat{\Sigma}^{-1} = (22.9439 \quad 70.0536 \quad 18.2104 \quad 60.7408 \quad 688.7211 \quad -26.1440)'$$

$$\hat{\mu}_3'\hat{\Sigma}^{-1} = (21.9707 \quad 70.7400 \quad 20.9248 \quad 56.5007 \quad 652.8811 \quad 28.3541)'$$

$$\hat{\mu}_4'\hat{\Sigma}^{-1} = (19.9934 \quad 69.2849 \quad 24.0749 \quad 47.9479 \quad 575.8803 \quad -26.8590)'$$

$$\frac{1}{2}\hat{\mu}_1'\hat{\Sigma}^{-1}\hat{\mu}_1 = 2.012.5, \quad \frac{1}{2}\hat{\mu}_2'\hat{\Sigma}^{-1}\hat{\mu}_2 = 1.918.4, \quad \frac{1}{2}\hat{\mu}_3'\hat{\Sigma}^{-1}\hat{\mu}_3 = 1.869.1,$$

$$\frac{1}{2}\hat{\mu}_4'\hat{\Sigma}^{-1}\hat{\mu}_4 = 1.817.8, \quad \frac{1}{2}\hat{\mu}_5'\hat{\Sigma}^{-1}\hat{\mu}_5 = 1.670.6$$

利用上述原理, 得到下列评价函数:

$$y_5(X) = \ln\left(\frac{1}{5}\right) - 2.012.5 + 24.517X_1 + 66.958X_2 + 13.29X_3 + 81.63X_4 + 728.65X_5 - 10.029X_6 \quad (5 \text{ 类})$$

$$y_4(X) = \ln\left(\frac{1}{5}\right) - 1.918.4 + 23.836X_1 + 67.68X_2 + 16.116X_3 + 67.274X_4 + 700.09X_5 - 21.12X_6 \quad (4 \text{ 类})$$

$$y_3(X) = \ln\left(\frac{1}{5}\right) - 1.869.1 + 22.944X_1 + 70.054X_2 + 18.21X_3 + 60.741X_4 + 688.72X_5 - 26.144X_6 \quad (3 \text{ 类})$$

$$y_2(X) = \ln\left(\frac{1}{5}\right) - 1.817.8 + 21.971X_1 + 70.74X_2 + 20.925X_3 + 56.501X_4 + 652.88X_5 - 28.354X_6 \quad (2 \text{ 类})$$

$$y_1(X) = \ln\left(\frac{1}{5}\right) - 1.670.6 + 19.993X_1 + 69.285X_2 + 24.075X_3 + 47.948X_4 + 575.88X_5 - 26.859X_6 \quad (1 \text{ 类})$$

由上述评价函数, 可求得各样品的判别函数值, 判别结果如下:

样品 1 的判别函数值为 $y_1 = -1.670.6 + 19.993 \times 1.0 + 69.285 \times 1.0 + 24.075 \times 1.0 + 47.948 \times 1.0 + 575.88 \times 1.0 - 26.859 \times 1.0 = 630.37$

第一个样本, 分别求得各个判别函数的值为:

$$y_1^{(1)} = 1.612\ 5 \times 10^4, \quad y_1^{(2)} = 1.693\ 4 \times 10^3, \quad y_1^{(3)} = 1.739\ 3 \times 10^3, \\ y_1^{(4)} = 1.765\ 7 \times 10^3, \quad y_1^{(5)} = 1.767\ 9 \times 10^3,$$

待判样本 2 和 3, 可类似进行判别, 得到表 6.4。

表 6.4 评价结果

| 判 | 1 | 75.58 | 10.77 | 28.7 | 0.88 | 0.84 | 1.27 | 1 |
|---|---|-------|-------|------|------|------|------|---|
| 样 | 2 | 85.54 | 7.12 | 25.3 | 1.67 | 1.15 | 1.37 | 3 |
| 本 | 3 | 91.57 | 5.39 | 23.1 | 2.11 | 1.13 | 1.58 | 4 |

例 6.4 某地区有 4 种主要树种, 它们对 3 种大气污染物的反应不同, 现测得 3 种大气污染物的浓度, 并求得判别函数, 并判定另外 3 个待判样本属于哪一类。

解 由表 6.5 可知, 4 种主要树种对 3 种大气污染物的反应不同, 现测得 3 种大气污染物的浓度, 并求得判别函数, 并判定另外 3 个待判样本属于哪一类。

表 6.5 三种大气污染物下的植物反应

| 组别 | 序号 | 叶色指数 x_1 | 植株生长指数 x_2 |
|-------------|----|------------|--------------|
| 第一组
遭受污染 | 1 | 4.3 | 15.7 |
| | 2 | 5.6 | 17.8 |
| | 3 | 4.7 | 16.9 |
| | 4 | 4.8 | 16.3 |
| | 5 | 5.3 | 17.2 |
| | 6 | 4.1 | 16.0 |
| | 7 | 4.0 | 15.8 |
| | 8 | 4.6 | 16.2 |

续表

| 组别 | 序号 | 叶色指数 x_1 | 植株生长指数 x_2 |
|----------------------------|----|------------|--------------|
| 第一组
遭受 SO_2 污染 | 1 | 9.6 | 19.6 |
| | 2 | 9.7 | 19.7 |
| | 3 | 9.7 | 19.7 |
| | 4 | 8.8 | 18.9 |
| | 5 | 8.5 | 19.6 |
| 第二组
遭受 HCl 污染 | 1 | 10.2 | 30.3 |
| | 2 | 11.3 | 28.7 |
| | 3 | 9.8 | 25.6 |
| | 4 | 7.2 | 27.6 |
| | 5 | 8.5 | 29.0 |
| | 6 | 9.6 | 30.0 |
| 待判样本 | 1 | 9.2 | 19.0 |
| | 2 | 4.8 | 15.3 |
| | 3 | 11.2 | 30.3 |

解 根据表中的数据, 得到:

$$\hat{\mu}_1 = (4.675\ 0\ 16.487\ 5)$$

$$\hat{\mu}_2 = (8.980\ 0\ 19.320\ 0)$$

$$\hat{\mu}_3 = (9.433\ 3\ 28.533\ 3)$$

$$\hat{\Sigma} = \begin{pmatrix} 0.819\ 8 & 0.355\ 8 \\ 0.355\ 8 & 1.231\ 9 \end{pmatrix}$$

利用 Bayes 原理, 建立评价函数为:

$$y_1(x) = \ln\left(\frac{1}{3}\right) - 108.571\ 8 - 0.015\ 3x_1 + 13.174\ 5x_2$$

$$y_2(x) = \ln\left(\frac{1}{3}\right) - 157.550\ 1 + 4.854\ 9x_1 + 14.052\ 9x_2$$

$$y_3(x) = \ln\left(\frac{1}{3}\right) - 326.390\ 5 + 1.842\ 0x_1 + 22.268\ 9x_2$$

将待判样本 x_1, x_2 代入上述三个判别函数, 计算函数值, 并比较函数值, 将待判样本归入函数值最大的那一类。将表 6.5 中待判样本 x_1, x_2 代入上述三个判别函数, 计算函数值, 并比较函数值, 将待判样本归入函数值最大的那一类。将表 6.5 中待判样本 x_1, x_2 代入上述三个判别函数, 计算函数值, 并比较函数值, 将待判样本归入函数值最大的那一类。

评价结果

| 待
判
样
本 | 样本编号 | x_1 | x_2 | 期望值 |
|------------------|------|-------|-------|-----|
| | 1 | 9.2 | 19.0 | 2 |
| | 2 | 4.8 | 15.3 | |
| | 3 | 11.2 | 30.3 | 3 |

[illegible]

样本数据表

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---------|---------|---------|---------|---------|---------|
| 1 | 9.875 3 | 0.462 3 | 1.820 5 | 0.072 9 | 0.001 8 | 0.000 8 |
| 1 | 8.641 2 | 0.037 0 | 2.464 2 | 0.066 7 | 0.001 2 | 0.000 8 |
| 1 | 9.804 5 | 1.476 4 | 0.528 8 | 0.060 9 | 0.001 9 | 0.000 9 |
| 1 | 8.525 7 | 1.787 3 | 0.173 7 | 0.052 9 | 0.001 6 | 0.000 0 |
| | 7.847 2 | 0.405 5 | 0.596 2 | 0.090 6 | 0.000 5 | 0.000 2 |
| | 6.022 9 | 3.493 6 | 3.000 0 | 0.305 8 | 0.002 0 | 0.004 7 |
| 2 | 6.699 0 | 2.837 3 | 3.000 0 | 0.446 2 | 0.002 0 | 0.003 1 |
| 2 | 6.304 0 | 3.344 3 | 3.000 0 | 0.443 3 | 0.002 0 | 0.001 1 |
| 2 | 7.021 9 | 2.759 0 | 3.000 0 | 0.441 1 | 0.002 0 | 0.003 3 |
| 2 | 7.064 2 | 2.857 8 | 3.000 0 | 0.256 6 | 0.002 0 | 0.001 8 |
| 3 | 5.193 4 | 5.364 4 | 3.302 8 | 0.770 8 | 0.002 5 | 0.003 8 |
| 3 | 5.378 4 | 5.720 0 | 3.853 7 | 0.796 8 | 0.003 5 | 0.004 6 |
| 3 | 5.821 6 | 5.289 8 | 3.818 0 | 0.830 1 | 0.003 0 | 0.002 2 |
| 3 | 5.341 2 | 5.068 2 | 3.727 1 | 0.654 6 | 0.004 5 | 0.003 3 |
| 3 | 5.370 4 | 5.405 5 | 3.546 6 | 0.722 4 | 0.004 1 | 0.003 5 |
| 4 | 4.589 6 | 9.827 4 | 5.045 2 | 1.440 1 | 0.005 9 | 0.004 9 |
| 4 | 3.542 9 | 7.009 3 | 5.751 5 | 1.368 7 | 0.005 7 | 0.001 0 |
| 4 | 4.787 8 | 6.796 6 | 4.597 4 | 1.330 7 | 0.006 4 | 0.002 9 |
| 4 | 3.129 6 | 9.953 3 | 5.165 6 | 1.211 7 | 0.007 6 | 0.002 3 |
| 4 | 3.865 8 | 6.903 8 | 5.159 6 | 1.380 2 | 0.007 6 | 0.003 6 |

续表

| | $DO(x_1)$ | 高锰酸盐指数(x_2) | $BOD_5(x_3)$ | $NH_3-N(x_4)$ | 挥发酚(x_5) | 铜(x_6) |
|---|-----------|-----------------|--------------|---------------|--------------|------------|
| 5 | 2.209 1 | 11.899 1 | 9.133 3 | 1.840 4 | 0.051 5 | 0.007 8 |
| 5 | 2.794 2 | 10.295 9 | 8.411 5 | 1.525 1 | 0.047 4 | 0.006 5 |
| 5 | 2.874 4 | 10.075 0 | 9.071 8 | 1.985 4 | 0.099 1 | 0.008 9 |
| 5 | 2.438 7 | 12.491 6 | 6.855 9 | 1.821 7 | 0.038 8 | 0.009 8 |
| 5 | 2.726 6 | 12.059 8 | 8.978 3 | 1.634 0 | 0.049 6 | 0.009 7 |

表 6-8 2000 年 4 月份各站点水环境监测指标实测值

| 各站点 | 指 标 | | | | | |
|-----|-----------|-----------------|--------------|---------------|--------------|------------|
| | $DO(x_1)$ | 高锰酸盐指数(x_2) | $BOD_5(x_3)$ | $NH_3-N(x_4)$ | 挥发酚(x_5) | 铜(x_6) |
| 紫岐花 | 7.80 | 1.80 | 1.10 | 0.08 | 0.000 | 0.000 0 |
| 望江楼 | 1.40 | 6.20 | 24.90 | 6.22 | 0.018 | 0.000 0 |
| 高 场 | 8.00 | 2.40 | 0.80 | 0.08 | 0.000 | 0.001 3 |
| 朱 沱 | 9.42 | 3.04 | 0.50 | 0.12 | 0.000 | 0.001 5 |
| 寸 滩 | 9.40 | 2.70 | 0.60 | 0.00 | 0.001 | 0.001 4 |
| 张家界 | 9.80 | 1.00 | 0.60 | 0.06 | 0.000 | 0.000 0 |
| 古 首 | 7.60 | 2.50 | 5.60 | 0.22 | 0.000 | 0.000 0 |
| 芷 江 | 6.60 | 1.50 | 0.90 | 0.32 | 0.000 | 0.000 0 |
| 坝 上 | 7.60 | 3.90 | 3.00 | 1.00 | 0.000 | 0.000 0 |
| 津 市 | 8.00 | 1.90 | 0.90 | 0.22 | 0.002 | 0.000 0 |
| 石 门 | 9.00 | 1.30 | 0.60 | 0.22 | 0.002 | 0.000 0 |
| 益 阳 | 8.30 | 1.70 | 0.20 | 0.09 | 0.000 | 0.000 0 |
| 湘 潭 | 9.10 | 2.50 | 2.00 | 0.30 | 0.000 | 0.000 0 |
| 株 洲 | 7.30 | 3.80 | 1.60 | 0.22 | 0.000 | 0.000 0 |
| 衡 阳 | 8.00 | 3.10 | 1.50 | 0.34 | 0.000 | 0.000 0 |
| 长 沙 | 8.10 | 2.30 | 0.50 | 0.17 | 0.000 | 0.002 7 |
| 吉 安 | 8.70 | 2.50 | 2.70 | 0.21 | 0.000 | 0.000 0 |
| 中 山 | 10.10 | 1.70 | 1.50 | 0.10 | 0.000 | 0.000 0 |

解 根据表中的数据,得到样本均值:

$$\hat{\mu}_1 = (8.938\ 8\ 0.833\ 7\ 1.116\ 7\ 0.068\ 8\ 0.001\ 4\ 0.000\ 5)$$

$$\hat{\mu}_2 = (6.622\ 4\ 3.058\ 4\ 3.000\ 0\ 0.378\ 6\ 0.002\ 0\ 0.002\ 7)$$

$$\hat{\mu}_3 = (5.421\ 0\ 5.369\ 6\ 3.649\ 6\ 0.754\ 9\ 0.003\ 5\ 0.003\ 5)$$

$\hat{\mu}$ (2.608 6 11.364 3 8.490 2 1.761 3 0.057 3 0.008 5)

协方差:

| | | | | | | |
|----------|---------|-----------|---------|-----------|-----------|-----------|
| | 0.319 0 | 0.095 4 | 0.013 7 | 0.005 9 | 0.000 7 | 0.000 1 |
| | 0.095 4 | 0.926 6 | 0.214 0 | - 0.008 0 | - 0.003 6 | 0.000 4 |
| Σ | 0.013 7 | - 0.214 0 | 0.418 5 | 0.003 5 | 0.002 4 | - 0.000 1 |
| | 0.005 9 | - 0.008 0 | 0.003 5 | 0.010 7 | 0.000 5 | 0.000 0 |
| | 0.000 7 | 0.003 6 | 0.002 4 | 0.000 5 | 0.000 1 | 0.000 0 |
| | 0.000 1 | 0.000 4 | 0.000 1 | 0.000 0 | 0.000 0 | 0.000 0 |

利用 Bayes 原理, 所建立的评价函数为:

$$y_1(x) = \ln \left(\frac{1}{x} \right) = -141.6957 + 30.7x_1 + 6.1x_2 + 4.5x_3 + 6.1x_4 - 102.2x_5 - 3.2097x_6$$

$$y_7(x) = \ln \left(\begin{matrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{matrix} \right) - 112.879 + 23.2x_1 + 8.3x_2 + 11.7x_3 + 46.4x_4 - 322.4x_5 - 1928.5x_6$$

2 683.3 x_4

$$y_1(x) = \ln\left(\frac{1}{5}\right) - 262.3979 + 16.5x_1 + 16.0x_2 + 21.9x_3 + 176.5x_4 - 811.0x_5 - 5185.7x_6$$

$$y_i(x) = \ln\left(\frac{1}{5}\right) \cdot 418.112 + 11.3x_1 + 21.4x_2 + 31.4x_3 + 193.0x_4 + 156.6x_5 + 1976.9x_6$$

[illegible]

表 6.9

评价结果

| 各站点 | 角, 别函数值 ($\times 10^3$) | | | | | 期望值 |
|-----|---------------------------|---------|---------|----------|---------|-----|
| | I | II | III | IV | V | |
| 攀枝花 | 0.112 8 | 0.098 0 | 0.055 7 | -0.068 6 | 0.243 2 | 1 |
| 望江楼 | 0.086 2 | 0.342 9 | 0.918 7 | 1.485 7 | 1.702 8 | 5 |
| 高 场 | 0.117 1 | 0.101 7 | 0.058 4 | -0.069 0 | 0.240 1 | 1 |

续表

| 各站点 | 判别函数值($\times 10^3$) | | | | | 期望值 |
|-----|------------------------|---------|---------|----------|----------|-----|
| | I | II | III | IV | V | |
| 朱 沱 | 0.162 9 | 0.137 9 | 0.092 7 | -0.035 9 | -0.212 5 | 1 |
| 寸 滩 | 0.160 2 | 0.130 1 | 0.078 3 | 0.060 9 | 0.210 3 | 1 |
| 张家界 | 0.167 1 | 0.131 0 | 0.076 8 | -0.063 0 | -0.257 4 | |
| 古 首 | 0.132 2 | 0.158 2 | 0.141 8 | 0.062 4 | 0.062 2 | 2 |
| 芷 江 | 0.071 7 | 0.076 5 | 0.048 2 | -0.055 2 | 0.223 2 | 2 |
| 坝 上 | 0.133 7 | 0.175 7 | 0.192 4 | 0.165 7 | 0.036 7 | 3 |
| 津 市 | 0.119 3 | 0.107 0 | 0.070 1 | -0.045 0 | 0.219 0 | 1 |
| 石 门 | 0.145 1 | 0.121 7 | 0.078 6 | -0.044 7 | -0.230 0 | 1 |
| 益 阳 | 0.123 6 | 0.098 7 | 0.051 7 | -0.079 9 | -0.266 1 | 1 |
| 湘潭 | 0.162 5 | 0.134 7 | 0.124 3 | 0.022 5 | -0.142 9 | 1 |
| 株洲 | 0.112 8 | 0.115 4 | 0.089 5 | 0.009 1 | -0.163 3 | 2 |
| 衡 阳 | 0.130 3 | 0.130 2 | 0.105 4 | 0.010 2 | -0.150 4 | 1 |
| 长 沙 | 0.114 3 | 0.101 1 | 0.039 5 | -0.066 9 | 0.236 0 | 1 |
| 吉 安 | 0.152 8 | 0.149 4 | 0.118 4 | 0.015 4 | -0.142 8 | 1 |
| 中 山 | 0.184 9 | 0.156 1 | 0.108 4 | -0.020 1 | 0.203 0 | 1 |
| 宣 城 | 0.154 4 | 0.120 8 | 0.066 9 | -0.073 8 | -0.263 9 | 1 |

【思考题 6】

1. 试述距离判别的基本步骤
2. 试述 Fisher 判别的基本步骤
3. 试述 Bayes 判别的基本步骤

的日平均浓度值,见表 6.10。

表 6.10 各监测点的监测值表 单位: mg/m^3

| 测点号 | 日平均值 | | |
|-----|---------------|---------------|-------|
| | SO_2 | NO_x | TSP |
| 1 | 0.006 0 | 0.022 | 0.327 |
| 2 | 0.009 0 | 0.018 | 0.190 |
| 3 | 0.010 5 | 0.015 | 0.254 |
| 4 | 0.004 0 | 0.025 | 0.160 |
| 5 | 0.003 0 | 0.018 | 0.053 |

函数。

表 6.11 大气环境质量标准表 (GB 3095—1996) 单位: mg/m^3

| 污染物名称 | 日平均浓度限值 | | |
|---------------|---------|------|------|
| | 一级标准 | 二级标准 | 三级标准 |
| SO_2 | 0.05 | 0.15 | 0.25 |
| NO_x | 0.10 | 0.10 | 0.15 |
| TSP | 0.12 | 0.30 | 0.50 |

判别函数

表 6.12 河流 As、Pb 的污染监测表

| 地点 | 序号 | As | | Pb | |
|------|----|----------------------|-----------------------|----------------------|-----------------------|
| | | $/(mg \cdot L^{-1})$ | $/(mg \cdot kg^{-1})$ | $/(mg \cdot L^{-1})$ | $/(mg \cdot kg^{-1})$ |
| 甲地 | 1 | 4.67 | 22.31 | 12.31 | 47.80 |
| | 2 | 4.63 | 28.82 | 16.18 | 62.55 |
| | 3 | 3.54 | 15.29 | 7.58 | 43.20 |
| 乙地 | 1 | 1.06 | 2.18 | 1.22 | 20.60 |
| | 2 | 0.80 | 3.85 | 4.06 | 47.10 |
| | 3 | 0.00 | 11.40 | 3.50 | 0.00 |
| | 4 | 2.42 | 3.66 | 2.14 | 15.00 |
| 样本 A | | 2.79 | 13.85 | 7.80 | 49.60 |
| 样本 B | | 2.40 | 7.90 | 4.30 | 33.20 |

6. 试述判别判别、Fisher 判别、Bayes 判别的区别与联系。

【参考文献】

1. 王惠德, 王惠德. 环境判别分析. 北京: 中国环境出版社, 2004.
2. 王惠德, 王惠德. 环境判别分析. 北京: 中国环境出版社, 2004.
3. 王惠德, 王惠德. 环境判别分析. 北京: 中国环境出版社, 2004.
4. 王惠德, 王惠德. 环境判别分析. 北京: 中国环境出版社, 2004.
5. 王惠德, 王惠德. 环境判别分析. 北京: 中国环境出版社, 2004.
6. 王惠德, 王惠德. 环境判别分析. 北京: 中国环境出版社, 2004.
7. 王惠德, 王惠德. 环境判别分析. 北京: 中国环境出版社, 2004.
8. 王惠德, 王惠德. 环境判别分析. 北京: 中国环境出版社, 2004.
9. 王惠德, 王惠德. 环境判别分析. 北京: 中国环境出版社, 2004.
10. 王惠德, 王惠德. 环境判别分析. 北京: 中国环境出版社, 2004.

环保, 2004, 30(5): 39-40.

第7章 环境主成分分析

主成分分析(Principal Component Analysis, PCA)是一种降维技术,它通过线性变换将原始数据转换为少数几个互不相关的主成分,从而简化数据结构,保留数据的主要信息。在环境科学中,PCA常用于处理多变量数据,如水质监测数据、气象数据等,以揭示数据中的潜在模式和主要影响因素。

本章将详细介绍PCA的基本原理、计算步骤及其在环境科学中的应用。首先,我们将回顾PCA的数学基础,包括协方差矩阵、特征值和特征向量的概念。然后,我们将通过具体的例子演示如何进行PCA分析,并讨论如何解释主成分的含义。最后,我们将探讨PCA在环境数据中的实际应用,如污染源识别、环境质量评价等。

本章的主要内容是:

- 主成分分析概述;
- 主成分分析计算原理;
- 主成分分析基本性质;
- 环境应用。

7.1 主成分分析概述

主成分分析(PCA)是一种多元统计方法,旨在通过线性变换将高维数据转换为低维表示,同时最大限度地保留数据的方差。在环境科学中,PCA常用于处理多变量数据,如水质监测数据、气象数据等,以揭示数据中的潜在模式和主要影响因素。PCA的核心思想是找到数据中方差最大的方向,即主成分。第一个主成分解释了数据中最大的方差,第二个主成分解释了剩余的方差中最大的部分,依此类推。通过选择前几个主成分,可以有效地降维,简化数据结构,便于后续分析和解释。PCA广泛应用于环境科学中的许多领域,如污染源识别、环境质量评价、气候变化研究等。通过PCA,研究人员可以识别出影响环境指标的主要因素,并评估不同因素对环境的贡献。此外,PCA还可以用于数据可视化,帮助研究人员直观地理解数据的分布和结构。

在对多变量数据进行主成分分析时,首先需要对数据进行标准化处理,以消除量纲的影响。然后,计算数据的协方差矩阵,并求解其特征值和特征向量。特征值的平方根即为标准差的平方,特征向量则代表了主成分的方向。通过选择前几个最大的特征值对应的特征向量,可以得到数据的主要成分。在实际应用中,PCA可以帮助我们识别出影响环境指标的主要因素,并评估不同因素对环境的贡献。例如,在水质监测中,PCA可以识别出影响水质指标的主要因素,如温度、pH值、溶解氧等,并评估它们对水质的影响程度。

主成分分析。主成分分析是多元统计中一种降维技术,它通过线性变换,将原来高维数据中的信息,集中到少数几个主成分上,从而达到降维的目的。主成分分析在环境科学中应用广泛,如环境监测、环境影响评价、环境质量管理等。主成分分析的优点是:能减少数据的冗余性,提高数据的可解释性;能降低数据的维度,减少计算量;能提高数据的稳定性,减少噪声的影响。主成分分析的缺点是:主成分的解释性较差,难以理解;主成分的计算较为复杂,需要一定的数学基础。

主成分分析的基本步骤如下:首先,对原始数据进行标准化处理,消除量纲的影响;其次,计算协方差矩阵,并求出其特征值和特征向量;然后,根据特征值的大小,选择前几个主成分;最后,将原始数据投影到主成分空间,得到降维后的数据。主成分分析的结果通常用主成分得分和主成分贡献率来描述。主成分得分反映了每个样本在主成分上的位置,主成分贡献率反映了每个主成分对总方差的贡献。主成分分析的结果可以用于聚类分析、判别分析、回归分析等多种统计方法。主成分分析在环境科学中的应用实例如下:在环境监测中,主成分分析可以用于识别主要污染源;在环境影响评价中,主成分分析可以用于评价不同方案的环境影响;在环境质量管理中,主成分分析可以用于监测环境质量的变化。

7.2 主成分分析计算原理

主成分分析的计算原理如下:假设原始数据矩阵为 X , 其大小为 $n \times p$, 其中 n 为样本数, p 为变量数。首先,对 X 进行标准化处理,得到标准化数据矩阵 Z 。然后,计算 Z 的协方差矩阵 S , 其大小为 $p \times p$ 。接着,求出 S 的特征值和特征向量。特征值的大小反映了主成分的方差,特征向量反映了主成分的方向。根据特征值的大小,选择前 k 个主成分,记为主成分得分矩阵 F 。最后,将原始数据 X 投影到主成分空间,得到降维后的数据 F 。

主成分分析的计算原理可以用数学公式表示如下:假设原始数据矩阵为 X , 其大小为 $n \times p$, 其中 n 为样本数, p 为变量数。首先,对 X 进行标准化处理,得到标准化数据矩阵 Z 。然后,计算 Z 的协方差矩阵 S , 其大小为 $p \times p$ 。接着,求出 S 的特征值和特征向量。特征值的大小反映了主成分的方差,特征向量反映了主成分的方向。根据特征值的大小,选择前 k 个主成分,记为主成分得分矩阵 F 。最后,将原始数据 X 投影到主成分空间,得到降维后的数据 F 。主成分分析的计算原理可以用数学公式表示如下:假设原始数据矩阵为 X , 其大小为 $n \times p$, 其中 n 为样本数, p 为变量数。首先,对 X 进行标准化处理,得到标准化数据矩阵 Z 。然后,计算 Z 的协方差矩阵 S , 其大小为 $p \times p$ 。接着,求出 S 的特征值和特征向量。特征值的大小反映了主成分的方差,特征向量反映了主成分的方向。根据特征值的大小,选择前 k 个主成分,记为主成分得分矩阵 F 。最后,将原始数据 X 投影到主成分空间,得到降维后的数据 F 。

$F_1 = (a_{11}, a_{12}, \dots, a_{1p})'$, $F_2 = (a_{21}, a_{22}, \dots, a_{2p})'$, ..., $F_p = (a_{p1}, a_{p2}, \dots, a_{pp})'$, $\mathbf{x} = (x_1, x_2, \dots, x_p)'$, 也即:

$$F_1 = \mathbf{a}'_1 \mathbf{x}, \|\mathbf{a}_1\| = 1$$

为了使 F_1 与 F_2, \dots, F_p 不相关, 即 F_1 与 F_2, \dots, F_p 的协方差为零, 我们令 F_1 与 F_2, \dots, F_p 的协方差矩阵为零, 即

$$F_1 F_2' = (a_1' \mathbf{X} \mathbf{X} a_2 = a_1' \mathbf{X} a_2 = 0$$

从而, $\mathbf{V} = \frac{1}{n} \mathbf{X} \mathbf{X}' = \mathbf{V} a_1 a_1' + \dots + \mathbf{V} a_p a_p'$ 中 $\mathbf{V} a_1 a_1'$ 与 $\mathbf{V} a_2 a_2'$ 等项正交, 此时, \mathbf{V} 就是 \mathbf{X} 的相关系数矩阵。

把上面的问题写成数学表达式, 即求优化问题:

$$\max_{a_1' a_1 = 1} \mathbf{a}'_1 \mathbf{V} \mathbf{a}_1$$

利用拉格朗日乘数法, 构造拉格朗日函数 L 如下:

$$L = \mathbf{a}'_1 \mathbf{V} \mathbf{a}_1 - \lambda_1 (\mathbf{a}'_1 \mathbf{a}_1 - 1)$$

对 L 分别求关于 \mathbf{a}_1 和 λ_1 的偏导, 并令其为零, 有:

$$\frac{\partial L}{\partial \mathbf{a}_1} = 2\mathbf{V} \mathbf{a}_1 - 2\lambda_1 \mathbf{a}_1 = 0 \quad (7.1)$$

$$\frac{\partial L}{\partial \lambda_1} = -(\mathbf{a}'_1 \mathbf{a}_1 - 1) = 0 \quad (7.2)$$

由式(7.1)得:

$$\mathbf{V} \mathbf{a}_1 = \lambda_1 \mathbf{a}_1 \quad (7.3)$$

将式(7.3)代入式(7.2)中, 可求得 λ_1 的表达式, 代入式(7.3)中, 可得 \mathbf{a}_1 的表达式。根据目标函数及上式, 有:

$$D(F_1) = \mathbf{a}'_1 \mathbf{V} \mathbf{a}_1 = \mathbf{a}'_1 (\lambda_1 \mathbf{a}_1) = \lambda_1 \mathbf{a}'_1 \mathbf{a}_1 = \lambda_1 \quad (7.4)$$

所以, \mathbf{a}_1 所对应的特征值 λ_1 应取到最大值。

同理可得, 使 $F_2 = (a_{21}, a_{22}, \dots, a_{2p})'$ 与 F_1 及 F_3, \dots, F_p 不相关的单位特征向量 \mathbf{a}_2 中, \mathbf{a}_2 所对应的特征值 $\lambda_2 = \mathbf{a}'_2 \mathbf{V} \mathbf{a}_2$ 应取到最大值。

接着, 可以求得与 $\mathbf{a}_1, \mathbf{a}_2$ 正交的 \mathbf{a}_3 , 使 $\mathbf{a}'_3 \mathbf{a}_1 = \mathbf{a}'_3 \mathbf{a}_2 = 0$, 且使 F_3 在第一主成分上, 第二主成分上, 第三主成分上, 第四主成分上的方差为:

$$D(F_3) = \frac{1}{n} \mathbf{a}'_3 \mathbf{X} \mathbf{X} a_3 = \mathbf{a}'_3 \mathbf{V} \mathbf{a}_3$$

写成优化问题, 即:

$$\max \mathbf{a}'_3 \mathbf{V} \mathbf{a}_3$$

$$\mathbf{a}'_2 \mathbf{a}_3 = 0, \mathbf{a}'_1 \mathbf{a}_3 = 0$$

$$\eta = \frac{\sum_{i=1}^m \lambda_i}{\sum_{i=1}^m \lambda_i} \quad (m \geq h)$$

(5)求第 h 主成分 F_h , 有

$$F_h = \mathbf{a}'_h \mathbf{x} = \sum_{j=1}^p a_{hj} x_j \quad (7.11)$$

于是, 主成分 F_1, F_2, \dots, F_h 的表达式为: $F_1 = a_{11}x_1 + a_{12}x_2 + \dots + a_{1p}x_p$, $F_2 = a_{21}x_1 + a_{22}x_2 + \dots + a_{2p}x_p$, \dots , $F_h = a_{h1}x_1 + a_{h2}x_2 + \dots + a_{hp}x_p$. 用 F_1, F_2, \dots, F_h 代替 x_1, x_2, \dots, x_p 个变量。

7.3 主成分分析的性质

主成分分析主要有以下几条基本性质:

(1)主成分 F_1, F_2, \dots, F_h 的方差 $D(F_i) = \lambda_i$ 为特征值 λ_i , 有

$$D(F_1) = \lambda_1, D(F_2) = \lambda_2, \dots, D(F_h) = \lambda_h, \dots, \sum_{i=1}^h \lambda_i = \sum_{i=1}^p \lambda_i = \sum_{i=1}^p \sigma_i^2$$

式中, $F_h(t)$ 是 F_h 的第 t 个分量。

(2) F_h 的样本方差等于 λ_h , 即:

$$D(F_h) = \lambda_h$$

这个结论在式(7.9)中已经给予证明。

(3)主成分 F_1, F_2, \dots, F_h 的样本协方差为

$$\text{Cov}(F_h, F_l) = 0 \quad (\forall l \neq h) \quad (7.12)$$

证明:

$$\text{Cov}(F_h, F_l) = \mathbf{a}'_h \mathbf{V} \mathbf{a}_l = \mathbf{a}'_h (\lambda_l \mathbf{a}_l) = \lambda_l \mathbf{a}'_h \mathbf{a}_l = 0$$

从上述证明可知, 主成分分析, 即将式(7.1)中的 p 个变量变换成 h 个相互独立的主成分 F_1, F_2, \dots, F_h 的过程, 可将式(7.1)中的 p 个变量, 变换成 h 个相互独立的主成分 F_1, F_2, \dots, F_h , 且 F_1, F_2, \dots, F_h 的方差分别为 $\lambda_1, \lambda_2, \dots, \lambda_h$, 且 F_1, F_2, \dots, F_h 的协方差为 0。因此, 主成分分析, 可将式(7.1)中的 p 个变量, 变换成 h 个相互独立的主成分 F_1, F_2, \dots, F_h , 且 F_1, F_2, \dots, F_h 的方差分别为 $\lambda_1, \lambda_2, \dots, \lambda_h$, 且 F_1, F_2, \dots, F_h 的协方差为 0。因此, 主成分分析, 可将式(7.1)中的 p 个变量, 变换成 h 个相互独立的主成分 F_1, F_2, \dots, F_h , 且 F_1, F_2, \dots, F_h 的方差分别为 $\lambda_1, \lambda_2, \dots, \lambda_h$, 且 F_1, F_2, \dots, F_h 的协方差为 0。

从上面的讨论可知, 主成分分析过程可示意为:

$$(x_1, \dots, x_p) \xrightarrow{\text{主成分分析}} (F_1, \dots, F_h) \quad (m < p)$$

其中 F_1, F_2, \dots, F_h 是 p 个变量 x_1, x_2, \dots, x_p 下的数据表, 它是 p 维数据表 $X = (x_1, x_2, \dots, x_p)$ 的 h 个主成分, 且 F_1, F_2, \dots, F_h 的方差分别为 $\lambda_1, \lambda_2, \dots, \lambda_h$, 且 F_1, F_2, \dots, F_h 的协方差为 0。

$$F = (F_1, \dots, F_m) = \begin{bmatrix} e'_1 \\ \vdots \\ e'_n \end{bmatrix}$$

\hat{e}_i 是数据表 F 的第 i 个样本点, 它为:

$$\hat{e}_i = (F_1(i), \dots, F_m(i))' \quad (i=1, 2, \dots, n)$$

因此, 数据表 F 的重心 \hat{g} 为: $\hat{g} = (\hat{g}_1, \dots, \hat{g}_m)$, 其中 $\hat{g}_k = \frac{1}{n} \sum_{i=1}^n F_k(i)$, 即

$$\hat{g} = \frac{1}{n} \sum_{i=1}^n \hat{e}_i = \begin{bmatrix} \frac{1}{n} \sum_{i=1}^n F_1(i) \\ \vdots \\ \frac{1}{n} \sum_{i=1}^n F_m(i) \end{bmatrix}$$

因此, \hat{e}_i 到重心 $\hat{g} = \mathbf{0}$ 的距离为:

$$d^2(\hat{e}_i, \mathbf{0}) = \sum_{k=1}^m F_k^2(i) \quad (7.13)$$

这里, \hat{e}_i 到重心 \hat{g} 的距离 $d(\hat{e}_i, \hat{g})$ 为 F 的第 i 个样本点 \hat{e}_i 在 F 的重心 \hat{g} 上的投影 \hat{g}_i 到重心 \hat{g} 的距离 $d(\hat{g}_i, \hat{g})$ 的平方。这里, $\hat{g}_i = (F_1(i), \dots, F_m(i))'$, $\hat{g} = (\hat{g}_1, \dots, \hat{g}_m)$, $F = (F_1, \dots, F_m)$ 是数据表, $\hat{e}_i = (F_1(i), \dots, F_m(i))'$ 是数据表 F 的第 i 个样本点, $\hat{g} = (\hat{g}_1, \dots, \hat{g}_m)$ 是数据表 F 的重心, $d(\hat{e}_i, \hat{g})$ 是样本点 \hat{e}_i 到重心 \hat{g} 的距离为:

$$d^2(\hat{e}_i, \mathbf{0}) = \sum_{k=1}^m F_k^2(i) \quad (7.14)$$

7.4 环境应用

环境统计是环境科学的重要组成部分, 环境统计是研究环境变化与各种因素有关的一门科学, 环境统计的主要任务是: 收集、整理、分析和解释环境数据。除此之外, 主成分还有下列一些用途(陈玉成等, 1998):

(1) 环境评价, 环境评价是环境统计的一个重要组成部分;

(2) 环境质量评价, 环境质量评价是环境统计的一个重要组成部分, 评价环境质量的空间分布规律;

(3) 环境预测, 环境预测是环境统计的一个重要组成部分;

(4) 分析环境污染的理化过程;

(5) 环境质量监测的优化布点;

环境统计的主要任务是: 收集、整理、分析和解释环境数据。环境统计, 从数据

据中提取更多的有用信息。

在进行主成分分析时,需注意以下几点:

(1) 主成分分析中,主成分个数一般不超过原变量个数,且主成分之间应相互独立,无相关性。因此,在提取主成分时,应尽量选择那些方差较大,且彼此之间相关性较小的变量。

(2) 主成分分析中,主成分的物理意义应尽量明确。如果主成分的物理意义不明确,则主成分分析的结果将难以解释。因此,在提取主成分时,应尽量选择那些物理意义明确的变量。

(3) 主成分分析中,主成分的个数应尽量少。因为主成分个数越多,解释力越强,但计算量也越大。因此,在提取主成分时,应尽量选择那些解释力较强的主成分。

例 7.1 某地区 1990 年 1 月 1 日至 12 月 31 日的大气环境数据如下:

表 7.1 给出了该地区 1990 年 1 月 1 日至 12 月 31 日的大气环境数据。表中数据为各季度中各评价指标的实测浓度值。试对各季度大气环境质量进行评价。

表 7.1 大气环境质量评价标准 单位: mg/m^3

| 大气环境
质量级别 | 指标 | | |
|----------------|--------------------|--------------------|-------------------|
| | $\text{SO}_2(x_1)$ | $\text{NO}_x(x_2)$ | $\text{TPS}(x_3)$ |
| I 级(e_1) | 0.05 | 0.05 | 0.15 |
| II 级(e_2) | 0.15 | 0.10 | 0.30 |
| III 级(e_3) | 0.25 | 0.15 | 0.50 |
| IV 级(e_4) | 0.85 | 0.50 | 1.70 |

注: IV 为严重污染临界浓度。

表 7.2 各季度单元中各评价指标实测浓度值 单位: mg/m^3

| 季度单元 | 指标 | | |
|---------------|--------------------|--------------------|-------------------|
| | $\text{SO}_2(x_1)$ | $\text{NO}_x(x_2)$ | $\text{TPS}(x_3)$ |
| 第一季度(e_5) | 0.046 | 0.036 | 0.086 |
| 第二季度(e_6) | 0.139 | 0.044 | 0.152 |
| 第三季度(e_7) | 0.032 | 0.014 | 0.159 |
| 第四季度(e_8) | 0.056 | 0.016 | 0.183 |

解 原决策矩阵:

$$X = \begin{bmatrix} 0.050 & 0 & 0.050 & 0 & 0.150 & 0 \\ 0.150 & 0 & 0.100 & 0 & 0.300 & 0 \\ 0.250 & 0 & 0.150 & 0 & 0.500 & 0 \\ 0.850 & 0 & 0.500 & 0 & 1.700 & 0 \\ 0.046 & 0 & 0.036 & 0 & 0.086 & 0 \\ 0.139 & 0 & 0.044 & 0 & 0.152 & 0 \\ 0.032 & 0 & 0.014 & 0 & 0.159 & 0 \\ 0.056 & 0 & 0.016 & 0 & 0.183 & 0 \end{bmatrix}$$

标准差标准化处理后的矩阵为:

$$A = \begin{bmatrix} 0.534 & 7 & -0.392 & 1 & -0.470 & 3 \\ 0.170 & 0 & -0.084 & 5 & -0.192 & 3 \\ 0.194 & 7 & 0.222 & 9 & 0.178 & 1 \\ 2.382 & 8 & 2.374 & 8 & 2.402 & 5 \\ -0.549 & 3 & -0.478 & 0 & 0.588 & 9 \\ -0.210 & 2 & -0.428 & 9 & 0.466 & 6 \\ -0.600 & 4 & 0.613 & 3 & -0.453 & 6 \\ 0.512 & 9 & 0.601 & 0 & -0.409 & 2 \end{bmatrix}$$

$$C = \begin{bmatrix} 1.000 & 0 & 0.993 & 6 & 0.992 & 5 \\ 0.993 & 6 & 1.000 & 0 & 0.993 & 1 \\ 0.992 & 5 & 0.993 & 1 & 1.000 & 0 \end{bmatrix}$$

标准正交的。

$$U = \begin{bmatrix} 0.333 & 3 & 0.377 & 1 & 0.355 & 1 \\ 0.333 & 4 & 0.122 & 9 & -0.499 & 9 \\ 0.333 & 3 & 0.500 & 1 & 0.145 & 0 \end{bmatrix}$$

$$\lambda_1 = 2.986 \ 2 \quad \lambda_2 = 0.007 \ 6 \quad \lambda_3 = 0.006 \ 3$$

4. 累计贡献率: $\alpha_1 = 0.995 \ 4 > 85\%$

5. 求第一主成分 F_1 , 有:

$$F_1 = Au$$

$$\begin{pmatrix} 0.165 & 7 & 0.149 & 0 & 0.198 & 6 & 2.386 & 7 & -0.538 & 8 & -0.368 & 5 \\ 0.555 & 8 & 0.507 & 7 \end{pmatrix}$$

川东地区我国天然气开采较早、储量较丰富的区域,在满足四川省和重庆、湖北、湖南用气的基础上,尚有大量剩余气量。川东地区孝泉、新场、合兴场气田及王场、白夹河等气田,位于成都平原西缘,在德阳以北,跨德阳市与绵竹市、什邡市、广汉市、成都青白江区等众多乡镇,开发距离约 80 km,地质条件为东缘,地质构造较简单,地质储量较丰富。川东地区地理位置见图 7-2,孝泉、合兴场、新场气田在地区图上及地质剖面图、地质构造图、剖面图。

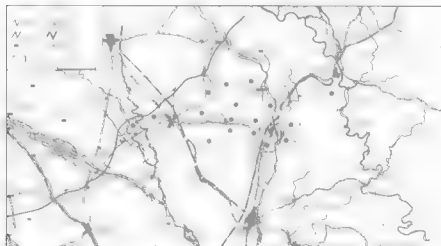


图 7-2 德阳地区孝泉、新场、合兴场气田开发工程范围示意图

在区域上选取各主要支流上选择多个具有代表性的地表水采样点,根据一般水质监测要求,主要监测指标有 pH 值、COD、氨氮等。将采集到的水质数据按照时间、地点、流量等进行分类,并分别进行统计分析,得出不同时间、地点、流量下的水质数据,并代表了一般地表水水质分布特征,具体见表 7-1。

表 7.5 天然气开发前、后流域地表水水质监测数据平均值和国标值

| | | 单位:mg/L.(pH 值除外) | | | |
|----------|-----|------------------|-------------------|-------|-------|
| | | pH 值 | COD _{Mn} | 石油类 | 挥发酚 |
| 国家 1 级标准 | | 9.00 | 15.00 | 0.050 | 0.002 |
| 国家 2 级标准 | | 9.00 | 20.00 | 0.050 | 0.005 |
| 国家 3 级标准 | | 9.00 | 30.00 | 0.500 | 0.010 |
| 开发前 | 绵远河 | 7.56 | 14.47 | 0.231 | 0.001 |
| | 凯江 | 9.25 | 28.89 | 0.125 | 0.001 |
| | 石亭江 | 8.08 | 22.48 | 0.120 | 0.001 |
| | 绵远河 | 8.16 | 10.37 | 0.071 | 0.001 |
| 开发后 | 凯江 | 7.31 | 20.07 | 0.059 | 0.001 |
| | 石亭江 | 7.41 | 31.04 | 0.103 | 0.001 |

解 由式 (7.1.1) 可得, 开发前、后流域地表水水质监测数据平均值和国标值矩阵 A_1 (开发前)、 A_2 (开发后)。

$$\begin{aligned}
 A_1 &= \begin{pmatrix} 0.5253 & -0.1234 & -0.7584 & -0.3689 \\ 0.5253 & -0.2716 & -0.7584 & 0.4611 \\ 0.5253 & 1.2319 & 1.8803 & 1.8443 \\ -1.6258 & -1.1031 & 0.3030 & -0.6455 \\ 0.8988 & 1.0650 & -0.3186 & -0.6455 \\ -0.8490 & 0.1012 & -0.3479 & -0.6455 \\ 0.8503 & -0.7454 & -0.4990 & -0.3689 \\ 0.8503 & 0.1324 & -0.4990 & 0.4611 \end{pmatrix} \\
 A_2 &= \begin{pmatrix} 0.8503 & 1.0936 & 2.0286 & 1.8443 \\ 0.1899 & -1.3131 & -0.3810 & -0.6455 \\ -1.2424 & 1.0650 & -0.3186 & -0.6455 \\ -0.8490 & 0.1012 & 0.3479 & -0.6455 \end{pmatrix}
 \end{aligned}$$

于是, 可得, 开发前、后流域地表水水质监测数据平均值和国标值矩阵 A_1 (开发前)、 A_2 (开发后)。

$$C_1 = \begin{bmatrix} 0.999\ 9 & 0.526\ 3 & 0.058\ 5 & 0.406\ 9 \\ 0.526\ 3 & 1.000\ 0 & 0.517\ 9 & 0.496\ 7 \\ -0.058\ 5 & 0.517\ 9 & 1.000\ 0 & 0.726\ 5 \\ 0.406\ 9 & 0.496\ 7 & 0.726\ 5 & 1.000\ 0 \end{bmatrix}$$

$$C_2 = \begin{bmatrix} 1.000\ 0 & -0.155\ 9 & 0.346\ 2 & 0.658\ 6 \\ -0.155\ 9 & 1.000\ 0 & 0.593\ 3 & 0.474\ 0 \\ 0.346\ 2 & 0.593\ 3 & 1.000\ 0 & 0.872\ 1 \\ 0.658\ 6 & 0.474\ 0 & 0.872\ 1 & 1.000\ 0 \end{bmatrix}$$

(3) 分别求其特征矩阵 U_1 、 U_2 和特征值。

$$U_1 = \begin{bmatrix} 0.464\ 4 & 0.256\ 9 & 0.767\ 0 & 0.360\ 6 \\ -0.341\ 0 & -0.742\ 8 & 0.200\ 5 & 0.540\ 7 \\ 0.631\ 2 & -0.084\ 9 & .588\ 1 & 0.498\ 5 \\ -0.519\ 9 & 0.612\ 4 & .160\ 1 & 0.573\ 0 \end{bmatrix}$$

$$U_2 = \begin{bmatrix} 0.390\ 9 & 0.451\ 0 & 0.716\ 4 & 0.361\ 4 \\ -0.157\ 6 & 0.617\ 1 & .668\ 3 & 0.384\ 5 \\ -0.469\ 1 & -0.643\ 8 & .146\ 5 & 0.586\ 5 \\ 0.776\ 1 & -0.036\ 8 & 0.136\ 6 & 0.614\ 5 \end{bmatrix}$$

表 7.6 特征值、贡献率及累计贡献率一览表

| 序号 | 项目开发前 | | | 项目开发后 | | |
|--------|-------------|-------------|-------------|-------------|-------------|-------------|
| | λ_1 | λ_2 | λ_3 | λ_1 | λ_2 | λ_3 |
| 特征值 | 2.061 6 | 0.728 9 | 0.209 4 | 2.285 0 | 0.567 6 | 0.147 5 |
| 贡献率% | 68.72 | 24.30 | 6.98 | 76.17 | 18.92 | 4.92 |
| 累计贡献率% | 68.72 | 93.02 | 100.00 | 76.17 | 95.09 | 100.00 |

根据式(7-11), F_1 和 F_2 的权重分别为 0.761 和 0.239 , 则 $F = \sum_{i=1}^2 F_i$ 。由表 7.6 可知, 项目开发前, 各主成分对总方差 F 的贡献率见表 7.6。从表中可以看到, 在项目进行之前, 有 68.72% 的方差是由 F_1 这一主成分解释, 而项目开发后, 有 76.17% 的

表 7.8 地表水环境质量标准

| 类别 | 总磷 (mg/L) | 氨氮 (mg/L) | 高锰酸盐指数 (mg/L) |
|-------|-----------|-----------|---------------|
| I 类 | 2 | 0.010 | 0.2 |
| II 类 | 4 | 0.025 | 0.5 |
| III 类 | 6 | 0.050 | 1.0 |
| IV 类 | 10 | 0.100 | 1.5 |
| V 类 | 15 | 0.200 | 2.0 |

表 7.9 2004 年太湖湖体主要污染指标浓度

| 湖区 | 高锰酸盐指数 | 总磷 ($\text{mg} \cdot \text{L}^{-1}$) | 总氮 ($\text{mg} \cdot \text{L}^{-1}$) |
|-------|--------|--|--|
| 五里湖 | 7.1 | 0.144 | 7.00 |
| 梅梁湖 | 5.7 | 0.102 | 5.27 |
| 西部沿岸区 | 5.4 | 0.107 | 3.33 |
| 东部沿岸区 | 4.0 | 0.056 | 1.71 |
| 湖心区 | 4.2 | 0.059 | 1.90 |
| 全湖平均 | 4.7 | 0.078 | 2.82 |

表 7.10 是太湖 10 个采样点土壤重金属元素测定结果, 用点因子法对采样点的 7 种重金属元素进行评价。

表 7.10 土壤重金属测定结果 单位: mg/km

| 序号 | Cd | Cr | Cu | Ni | Pb | Hg | As |
|----|-------|-------|-------|-------|-------|-------|-------|
| 1 | 0.221 | 89.52 | 42.66 | 32.63 | 46.50 | 0.530 | 10.90 |
| 2 | 0.462 | 57.21 | 46.49 | 25.42 | 27.35 | 0.082 | 7.82 |
| 3 | 0.132 | 73.28 | 31.40 | 34.38 | 37.98 | 0.370 | 11.47 |
| 4 | 0.109 | 57.88 | 25.70 | 25.82 | 31.11 | 0.114 | 7.54 |
| 5 | 0.078 | 44.57 | 36.60 | 22.06 | 22.65 | 0.187 | 7.39 |
| 6 | 0.129 | 63.34 | 22.63 | 26.85 | 23.86 | 0.033 | 6.90 |
| 7 | 0.132 | 74.83 | 18.57 | 31.71 | 32.54 | 0.137 | 9.08 |
| 8 | 0.170 | 73.32 | 56.27 | 41.84 | 27.45 | 0.746 | 10.46 |

续表

| 序号 | Cd | Cr | Cu | Ni | Pb | Hg | As |
|----|-------|-------|-------|-------|-------|-------|-------|
| 9 | 0.202 | 86.26 | 63.34 | 51.04 | 33.42 | 0.304 | 10.70 |
| 10 | 0.119 | 68.62 | 12.45 | 25.79 | 28.23 | 0.056 | 7.07 |
| 11 | 0.063 | 35.39 | 13.58 | 16.17 | 18.29 | 0.167 | 12.15 |
| 12 | 0.142 | 68.41 | 29.18 | 33.33 | 28.96 | 0.072 | 10.67 |
| 13 | 0.134 | 85.39 | 26.60 | 37.90 | 40.04 | 0.290 | 6.60 |
| 14 | 0.051 | 24.61 | 10.69 | 13.80 | 17.65 | 0.184 | 8.49 |
| 15 | 0.038 | 42.23 | 5.51 | 10.20 | 11.24 | 0.036 | 5.08 |
| 16 | 0.121 | 49.73 | 37.14 | 32.78 | 21.41 | 0.579 | 5.62 |
| 17 | 0.047 | 26.93 | 8.79 | 10.64 | 15.71 | 0.029 | 10.49 |
| 18 | 0.065 | 60.17 | 13.86 | 18.48 | 21.04 | 0.091 | 10.67 |
| 19 | 0.065 | 35.84 | 11.64 | 17.23 | 21.37 | 0.055 | 8.50 |
| 20 | 0.044 | 34.19 | 15.69 | 12.97 | 9.80 | 0.031 | 9.86 |
| 21 | 0.055 | 30.19 | 9.96 | 13.42 | 13.03 | 0.046 | 10.09 |
| 22 | 0.058 | 27.78 | 10.99 | 15.65 | 14.19 | 0.034 | 13.08 |

【参考文献】

7: 33-36.

2001, 30(4): 212-215.

第8章 环境因子分析

因子分析是一种多元统计方法，用于研究多个变量之间的内在结构。它通过提取少数几个公共因子来解释多个变量的变异。在环境科学中，因子分析常用于分析环境数据，如水质、大气污染等，以揭示潜在的环境因子及其对观测变量的影响。

结合环境案例分析该方法在环境科学中的应用。

本章的主要内容是：

- 因子分析概述；
- 正交因子模型；
- 正交因子模型的统计意义；
- 正交因子模型的求解；
- 因子旋转；
- 因子得分；
- 环境应用。

8.1 因子分析概述

因子分析是一种多元统计方法，用于研究多个变量之间的内在结构。它通过提取少数几个公共因子来解释多个变量的变异。在环境科学中，因子分析常用于分析环境数据，如水质、大气污染等，以揭示潜在的环境因子及其对观测变量的影响。

因子分析的基本思想是：假设观测到的多个变量是由少数几个不可观测的公共因子和独特的个别误差项共同作用的结果。通过数学方法，可以从观测变量中提取出这些公共因子，从而简化数据并揭示其内在结构。

在环境科学中，因子分析的应用非常广泛。例如，在水质分析中，可以通过因子分析识别出影响水质的主要因子，如无机磷、总磷等。在大气污染研究中，因子分析可以帮助识别出不同的污染源及其对空气质量的影响。

上式(8.1)中, x_i 为第 i 个指标的观测值, μ 为第 i 个指标的总平均, a_{ij} 为第 j 个因子对第 i 个指标的载荷, u_j 为第 j 个因子的随机误差。第(8.1)式可表示为矩阵形式, 其中 x 为 $1 \times n$ 行向量, μ 为 1×1 标量, A 为 $n \times p$ 矩阵, U 为 $1 \times p$ 行向量, n 为指标的个数, p 为因子的个数, i 为指标的序号, j 为因子的序号。记:

$$\begin{aligned} X &= (x_1, x_2, \dots, x_n)', A = (a_{ij})_{n \times p} \\ \mu &= (\mu_1, \mu_2, \dots, \mu_n)' \\ U &= (u_1, u_2, \dots, u_p)' \end{aligned}$$

则式(8.1)可以表示为:

$$X = \mu + AF + U$$

称其为因子模型。更一般地, 有下述定义:

设共有 n 个指标, p 个公共因子, $Z = (z_1, z_2, \dots, z_p)'$ 为公共因子向量, $\mu = (\mu_1, \mu_2, \dots, \mu_n)'$ 为指标的总平均向量, Σ 为指标的协方差矩阵, $\Sigma = \text{Cov}(X) = \text{Cov}(\mu + AF + U) = \text{Cov}(AF + U)$, 其中 $A = (a_{ij})_{n \times p}$ 为因子载荷矩阵, $F = (f_1, f_2, \dots, f_p)'$ 为公共因子向量, $U = (u_1, u_2, \dots, u_p)'$ 为随机误差向量, $\text{Cov}(U) = \Psi = \text{diag}(\psi_1^2, \psi_2^2, \dots, \psi_p^2)$ 为随机误差的协方差矩阵, $\text{Cov}(F) = I_p$ 为公共因子的协方差矩阵, I_p 为 p 阶单位矩阵, $\text{Cov}(F, U) = 0$ 为公共因子与随机误差的协方差矩阵, $\text{Cov}(F) = I_p$ 为公共因子的协方差矩阵, $\text{Cov}(U) = \Psi$ 为随机误差的协方差矩阵, $\text{Cov}(F, U) = 0$ 为公共因子与随机误差的协方差矩阵。

$$x_i = a_{i1}f_1 + a_{i2}f_2 + a_{i3}f_3 + \dots + a_{ip}f_p + \varepsilon_i \quad (8.2)$$

则式(8.2)可以表示为:

$$X = \mu + AF + U \quad (8.3)$$

其中, $X = (x_1, x_2, \dots, x_n)'$ 为指标的观测值向量, $\mu = (\mu_1, \mu_2, \dots, \mu_n)'$ 为指标的总平均向量, $A = (a_{ij})_{n \times p}$ 为因子载荷矩阵, $F = (f_1, f_2, \dots, f_p)'$ 为公共因子向量, $U = (u_1, u_2, \dots, u_p)'$ 为随机误差向量, $\text{Cov}(U) = \Psi = \text{diag}(\psi_1^2, \psi_2^2, \dots, \psi_p^2)$ 为随机误差的协方差矩阵, $\text{Cov}(F) = I_p$ 为公共因子的协方差矩阵, $\text{Cov}(F, U) = 0$ 为公共因子与随机误差的协方差矩阵, A 为因子载荷矩阵。

当 F 是随机向量时, 通常假定

$$\begin{aligned} E(F) &= 0, \text{Cov}(F) = I_p \\ E(U) &= 0, \text{Cov}(U) = \text{diag}(\psi_1^2, \dots, \psi_p^2) = \Psi \\ \text{Cov}(F, U) &= 0 \end{aligned} \quad (8.4)$$

式(8.4)中, $E(F) = 0$ 为公共因子的期望向量, $\text{Cov}(F) = I_p$ 为公共因子的协方差矩阵, $E(U) = 0$ 为随机误差的期望向量, $\text{Cov}(U) = \Psi$ 为随机误差的协方差矩阵, $\text{Cov}(F, U) = 0$ 为公共因子与随机误差的协方差矩阵。

公因子的依赖程度。

由式(8.10)可知, λ_i 为第 i 个因子 f_i 对变量 x_j 的方差贡献, 故 λ_i 为公因子 f_i 重要性的一个度量。

由式(8.10)可知, λ_i 为第 i 个因子 f_i 的方差贡献, 它的大小为因子 f_i 对变量 x_j 的含义提供了一种依据。

8.4 正交因子模型的求解

由式(8.10)可知, 正交因子模型 $X = AF + \Psi$ 中, 因子 F 的方差贡献为 λ_i , 故式(8.10)可写为

$$X = AF + \Psi \quad (8.4)$$

使其满足方差结构(8.5), 即满足:

$$\Sigma = AA' + \Psi$$

由式(8.4)可知, 正交因子模型 $X = AF + \Psi$ 中, 因子 F 的方差贡献为 λ_i , 故式(8.10)可写为

$$X = AF + \Psi \quad (8.4)$$

使其满足方差结构(8.5), 即满足:

由式(8.4)可知, 正交因子模型 $X = AF + \Psi$ 中, 因子 F 的方差贡献为 λ_i , 故式(8.10)可写为

$$X = AF + \Psi \quad (8.4)$$

使其满足方差结构(8.5), 即满足:

$$\begin{aligned} \Sigma &= PD_k P' = \lambda_1 e_1 e_1' + \lambda_2 e_2 e_2' + \cdots + \lambda_p e_p e_p' \\ &= (\sqrt{\lambda_1} e_1, \sqrt{\lambda_2} e_2, \cdots, \sqrt{\lambda_p} e_p) (\sqrt{\lambda_1} e_1, \sqrt{\lambda_2} e_2, \cdots, \sqrt{\lambda_p} e_p)' \\ &\quad - AA' \end{aligned} \quad (8.12)$$

由式(8.12)可知, 正交因子模型 $X = AF + \Psi$ 中, 因子 F 的方差贡献为 λ_i , 故式(8.10)可写为

$$X = AF + \Psi \quad (8.4)$$

使其满足方差结构(8.5), 即满足:

$$\Sigma = AA' + \Psi = AA' \quad (8.13)$$

由式(8.13)可知, 正交因子模型 $X = AF + \Psi$ 中, 因子 F 的方差贡献为 λ_i , 故式(8.10)可写为

$$X = AF + \Psi \quad (8.4)$$

使其满足方差结构(8.5), 即满足:

$$\Sigma \approx (\sqrt{\lambda_1} e_1, \sqrt{\lambda_2} e_2, \cdots, \sqrt{\lambda_k} e_k) (\sqrt{\lambda_1} e_1, \sqrt{\lambda_2} e_2, \cdots, \sqrt{\lambda_k} e_k)' - AA' \quad (8.14)$$

由式(8.14)可知, 正交因子模型 $X = AF + \Psi$ 中, 因子 F 的方差贡献为 λ_i , 故式(8.10)可写为

$$X = AF + \Psi \quad (8.4)$$

使其满足方差结构(8.5), 即满足:

$\Sigma = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix}$, $\Psi = \begin{pmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{pmatrix}$ 式中:

$\psi = (\psi_1, \psi_2, \dots, \psi_p)'$, $\psi_i = \sum_{j=1}^p \lambda_j$, 因此载荷矩阵 Λ 的估计为 $\Lambda = (\sqrt{\lambda_1}e_1, \sqrt{\lambda_2}e_2, \dots, \sqrt{\lambda_p}e_p)$ 。

由式 (8.14) $\Sigma = \Lambda\Lambda'$ 可知, 载荷矩阵 Λ 的估计为 $\Lambda = (\sqrt{\lambda_1}e_1, \sqrt{\lambda_2}e_2, \dots, \sqrt{\lambda_p}e_p)$ 。其中, e_1, e_2, \dots, e_p 为 Σ 的特征向量, 即 $\Sigma e_i = \lambda_i e_i$ ($i=1, 2, \dots, p$)。这里, e_i 可经下列计算求得 (在本办法中就是主成分函数), $i=1, 2, \dots, p$ 。由式 (8.14) 可知, 若用 R 表示式 (8.15) 的表示:

$$R = \Lambda\Lambda' + \Psi \\ = (\sqrt{\lambda_1}\hat{e}_1, \sqrt{\lambda_2}\hat{e}_2, \dots, \sqrt{\lambda_p}\hat{e}_p)(\sqrt{\lambda_1}\hat{e}_1, \sqrt{\lambda_2}\hat{e}_2, \dots, \sqrt{\lambda_p}\hat{e}_p)' + \text{diag}(\hat{\psi}_1^2, \hat{\psi}_2^2, \dots, \hat{\psi}_p^2) \quad (8.16)$$

由式 (8.16) 可知, $\hat{e}_1, \hat{e}_2, \dots, \hat{e}_p$ 为矩阵 R 的 p 个较大的特征值对应的特征向量, 且 $\hat{\psi}_1^2, \hat{\psi}_2^2, \dots, \hat{\psi}_p^2$ 为 $R = \Lambda\Lambda' + \Psi$ 的对角线元素, 由 R 出发, 因子分析模型的载荷矩阵的估计为:

$$\hat{A} = (\sqrt{\lambda_1}\hat{e}_1, \sqrt{\lambda_2}\hat{e}_2, \dots, \sqrt{\lambda_p}\hat{e}_p). \quad (8.17)$$

特殊因子的方差 $\hat{\psi}_i^2$ 的估计为:

$$\hat{\psi}_i^2 = 1 - \sum_{j=1}^p \hat{a}_{ij}^2 \quad (i=1, 2, \dots, p) \quad (8.18)$$

其中, \hat{a}_{ij} 为 \hat{A} 的 (i, j) 元素。

由式 (8.18) 可知, $\hat{\psi}_i^2$ 为特殊因子的方差, 因此可用 $\hat{\psi}_i^2$ 表示特殊因子的方差, 即 $\hat{\psi}_i^2 = \sigma^2(\epsilon_i)$ 。这里, ϵ_i 表示在多元正态分布的情况下, 特殊因子 ϵ_i 与 X 的协方差矩阵 R 的 i 行 i 列元素 R_{ii} 减去 $\sum_{j=1}^p \hat{a}_{ij}^2$ 的平方根, 即:

$$\hat{\sigma}_{\epsilon_i}^2 = \sum_{j=1}^p \hat{a}_{ij}^2 + (\sqrt{\lambda_j}\hat{e}_j)'(\sqrt{\lambda_j}\hat{e}_j) - \lambda_j \quad (8.19)$$

由此寻找一个 k 使得:

$$\left(\sum_{i=1}^k \hat{\lambda}_i / p \right) \times 100\% \geq 80\% (\text{或 } 75\%) \quad (8.20)$$

就确定该 k 为公因子数。

例 8.2 求例 8.1 中 5 个变量的 3 个公共因子, 并求计算相关矩阵的估计值 (估计值 \hat{R})。已知 5 个变量的累积方差贡献率见表 8.1。

表 8.1 R 的特征值及其累计方差贡献率

| 特征值 | 2.605 4 | 1.971 1 | 0.453 3 | 0.437 5 | 0.275 6 | 0.257 1 |
|-----------|---------|---------|---------|---------|---------|---------|
| 累计方差贡献率/% | 43.42 | 76.27 | 83.83 | 91.12 | 95.71 | 100.00 |

第 1 个因子方差贡献率相对于总方差占主导地位, 因此, 第 1 个因子从总体中独立出来, 用基式 (8.15) 表示, 我们称它为第 1 个主成分, 记为 f_1 。

根据式 (8.17) 得到:

$$A = \begin{pmatrix} 0.638 4 & 0.644 4 \\ 0.686 6 & -0.547 5 \\ 0.651 0 & 0.520 1 \\ 0.654 2 & 0.516 5 \\ 0.683 6 & 0.550 8 \\ 0.638 3 & 0.644 5 \end{pmatrix}$$

各公因子 f_j 对变量 x_i 的贡献 h_{ij}^2 , 即第 j 个变量的可解释方差:

$$\begin{aligned} h_{11}^2 &= |0.822 8| \\ h_{21}^2 &= |0.771 1| \\ h_{31}^2 &= |0.694 8| \\ h_{41}^2 &= |0.770 7| \\ h_{51}^2 &= |0.822 8| \end{aligned}$$

第 j 个公因子 f_j 对所有变量的贡献 R_j 分别为:

$$(g_1, g_2) = (2.605 4, 1.971 1)$$

由式 (8.15) 和 (8.18) 得:

$$\Psi = \begin{pmatrix} 0.177 2 & & & & & 0.000 0 \\ & 0.228 9 & & & & \\ & & 0.305 7 & & & \\ & & & 0.305 2 & & \\ & & & & 0.229 3 & \\ 0.000 0 & & & & & 0.177 2 \end{pmatrix}$$



图 8-1 坐标轴旋转图

$$A^* = AP = (a_{ij}^*) \quad (8.24)$$

$$d_{ij} = a_{ij}^* / h_i \quad (j=1, 2, \dots, k) \quad (8.25)$$

选择正交矩阵 P 使得 $\varphi = \frac{1}{p} \sum_{j=1}^k \sum_{i=1}^p \left(d_{ij}^2 - \frac{1}{p} \sum_{i=1}^p d_{ij}^2 \right)^2$ 达到最大(卢崇飞等, 1988)。

当 $k=2$ 时可以准确地求出 P 。

P 可以表示成:

$$P = \begin{pmatrix} \cos \theta & \sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \quad (8.27)$$

其中 θ 为 x' 轴与 x 轴的夹角。

$$\tan 4\theta = \frac{D}{C} \frac{2HG}{1/H^2 - G^2} = \frac{\alpha}{\beta} \quad (8.28)$$

其中:

$$D = 2 \sum_{i=1}^p B_i V_i, \quad H = \sum_{i=1}^p B_i, \quad G = \sum_{i=1}^p V_i,$$

$$C = \sum_{i=1}^n B_i^2 = \lambda_1^2 + \lambda_2^2 + \lambda_3^2 + \lambda_4^2 + \lambda_5^2 + \lambda_6^2 + \lambda_7^2 + \lambda_8^2 + \lambda_9^2 + \lambda_{10}^2 + \lambda_{11}^2 + \lambda_{12}^2 + \lambda_{13}^2 + \lambda_{14}^2 + \lambda_{15}^2 + \lambda_{16}^2 + \lambda_{17}^2 + \lambda_{18}^2 + \lambda_{19}^2 + \lambda_{20}^2$$

由式(8.26)求出的角 θ ，取 $\theta \in (0, \pi)$ ，使得 $\pi - \theta < \pi$ ，按此角 θ 作正交旋转，便可以使 φ 达到最大。

由式(8.26)求出的角 θ ，取 $\theta \in (0, \pi)$ ，使得 $\pi - \theta < \pi$ ，按此角 θ 作正交旋转，便可以使 φ 达到最大。

当 $a > 0$, $4\theta \in (0, \pi)$, $\theta \in (0, \frac{\pi}{4})$

当 $a < 0$, $4\theta \in (-\pi, 0)$, $\theta \in (-\frac{\pi}{4}, 0)$

对于 θ 的取值，我们采用以下方法：首先对每一因子 X_i 与因子 X_j 的载荷 A_{ij} 进行正交旋转，使得 A_{ij} 达到最大值，然后对每一因子 X_i 与因子 X_j 的载荷 A_{ij} 进行正交旋转，使得 A_{ij} 达到最大值，最后对每一因子 X_i 与因子 X_j 的载荷 A_{ij} 进行正交旋转，使得 A_{ij} 达到最大值。

由式(8.26)求出的角 θ ，取 $\theta \in (0, \pi)$ ，使得 $\pi - \theta < \pi$ ，按此角 θ 作正交旋转，便可以使 φ 达到最大。

由式(8.26)求出的角 θ ，取 $\theta \in (0, \pi)$ ，使得 $\pi - \theta < \pi$ ，按此角 θ 作正交旋转，便可以使 φ 达到最大。

例 8.3 我们利用例 8.1 的数据作为因子，将式(8.26)求出的角 θ 作正交旋转，便可以使 φ 达到最大。

未对载荷矩阵 A 进行因子旋转之前，方差

$$\varphi = 0.0060$$

通过式(8.27)、(8.28)和(8.29)，得

$$\theta = 0.7853$$

即： $\tan 4\theta = 2.5786 \times 10^{-4}$

$$P = \begin{pmatrix} 0.7072 & 0.7071 \\ 0.7071 & 0.7072 \end{pmatrix}$$

因此，将式(8.26)求出的角 θ 作正交旋转，便可以使 φ 达到最大。

$$A = AP$$

$$\begin{pmatrix} -0.004\ 2 & -0.907\ 1 \\ 0.098\ 4 & 0.872\ 6 \\ 0.092\ 6 & -0.828\ 1 \\ 0.827\ 8 & -0.097\ 3 \\ 0.872\ 9 & 0.093\ 8 \\ 0.907\ 1 & 0.004\ 5 \end{pmatrix}$$

载荷矩阵为载荷表, 如表 8-10 所示。因子旋转后, 为 2 变大

从表 8-10 载荷矩阵 A 可知, 因子载荷 f_1 和 f_2 分别含因子 SP 、 SO 和 NO 以及因子 CO 、 BC 、 CH 和 NI 。表示的大气环境因子 f_1 和 f_2 分别含因子 CO 、 BC 、 CH 和 NI 。表示的水环境因子。

因子旋转后, 各公共因子 f_1 和 f_2 的方差仍为 1, 为第一个变量的共同度不变, 仍为:

$$\begin{aligned} h_1^2 &= 0.822\ 8 \\ h_2^2 &= 0.771\ 1 \\ r_{12}^2 &= 0.694\ 3 \\ h_4^2 &= 0.694\ 8 \\ h_5^2 &= 0.770\ 7 \\ h_6^2 &= 0.822\ 8 \end{aligned}$$

因子旋转后, 第 1 个公共因子 f_1 和 2 个公共因子 f_2 的方差分别为

$$(g_1, g_2) = (2.288\ 3, 2.288\ 2)$$

因子旋转后的本征值 ψ 矩阵为下表所示。

$$\psi = \begin{pmatrix} 0.177\ 2 & & & & & 0.000\ 0 \\ & 0.228\ 9 & & & & \\ & & 0.305\ 7 & & & \\ & & & 0.305\ 2 & & \\ & & & & 0.229\ 3 & \\ 0.000\ 0 & & & & & 0.177\ 2 \end{pmatrix}$$

8.6 因子得分

到目前为止, 我们已会计算因子载荷矩阵 Σ 或者相关矩阵 R 并求得公共因子和因子载荷, 但如何由因子载荷矩阵确定公共因子含义。

因子得分, 如表 8-11 所示, 我们再次过来考察每一个公共因子, 分别各区域

8.7 环境应用

例 8.4 某地区 16 个大气颗粒物样品, 经分析得到 11 种元素含量, 见表 8.2。试利用主成分分析法, 对 16 个样品进行综合评价, 并得出相关结论。

表 8.2 样本中大气颗粒物成分分析结果表 单位: mg/kg

| 序号 | Br | K | Ba | Rb | Sc | Fe | Zn | Ni | V | W | As |
|----|-----|--------|-----|----|------|--------|-------|-----|-----|------|-----|
| 1 | 180 | 11 000 | 820 | 58 | 18.0 | 22 000 | 950 | 110 | 274 | 5.9 | 60 |
| 2 | 97 | 7 800 | 650 | 39 | 9.6 | 16 000 | 930 | 44 | 100 | 6.3 | 100 |
| 3 | 120 | 8 600 | 490 | 45 | 8.2 | 14 000 | 820 | 45 | 107 | 3.3 | 72 |
| 4 | 200 | 7 400 | 390 | 31 | 9.5 | 13 000 | 1 500 | 55 | 183 | 10.0 | 75 |
| 5 | 29 | 5 400 | 250 | 33 | 5.6 | 10 000 | 170 | 30 | 88 | 3.2 | 25 |
| 6 | 42 | 9 100 | 490 | 43 | 6.1 | 14 000 | 370 | 17 | 93 | 2.5 | 39 |
| 7 | 60 | 12 000 | 520 | 54 | 10.0 | 21 000 | 780 | 45 | 129 | 4.3 | 49 |
| 8 | 38 | 8 700 | 430 | 41 | 8.2 | 16 000 | 680 | 37 | 96 | 4.9 | 56 |
| 9 | 110 | 5 400 | 250 | 30 | 4.6 | 7 300 | 860 | 39 | 1 | 2.7 | 53 |
| 10 | 38 | 4 900 | 174 | 20 | 3.5 | 6 700 | 480 | 36 | 50 | 3.1 | 39 |
| 11 | 100 | 7 100 | 360 | 29 | 5.5 | 11 000 | 960 | 22 | 28 | 5.3 | 25 |
| 12 | 60 | 4 200 | 130 | 15 | 2.1 | 4 400 | 840 | 17 | 24 | 3.9 | 25 |
| 13 | 15 | 5 800 | 240 | 27 | 5.5 | 11 000 | 650 | 25 | 49 | 4.9 | 40 |
| 14 | 17 | 8 000 | 260 | 35 | 5.1 | 12 000 | 370 | 20 | 48 | 3.5 | 30 |
| 15 | 19 | 870 | 290 | 38 | 5.8 | 14 000 | 800 | 26 | 40 | 6.1 | 25 |
| 16 | 13 | 46 000 | 20 | 20 | 3.7 | 7 200 | 370 | 14 | 44 | 3.7 | 25 |

解 由表 8.2 数据, 建立数据矩阵, 然后进行标准化, 计算协方差矩阵 Σ 表示。

| | | | | | | | | | | |
|-------|-------|-------|-------|-------|--------|-------|--------|-------|--------|--------|
| 1.865 | 0.147 | 2.266 | 1.950 | 2.949 | 1.946 | 0.713 | 3.191 | 2.777 | 0.690 | 0.626 |
| 0.443 | 0.170 | 1.428 | 0.348 | 0.710 | 0.720 | 0.651 | 0.331 | 0.225 | 0.903 | 2.429 |
| 0.837 | 0.091 | 0.640 | 0.851 | 0.337 | 0.312 | 0.300 | 0.374 | 0.328 | 0.690 | 1.166 |
| 2.208 | 0.210 | 1.147 | 0.327 | 0.683 | 0.107 | 2.421 | 0.807 | 1.442 | 2.868 | 1.302 |
| 0.722 | 0.408 | 1.543 | 0.158 | 0.357 | 0.506 | 1.712 | -0.276 | 0.054 | -0.744 | 1.952 |
| 0.499 | 0.041 | 1.640 | 0.685 | 0.223 | 0.312 | 1.090 | 0.840 | 0.123 | -1.115 | 0.321 |
| 0.191 | 0.246 | 1.787 | 1.513 | 1.816 | 1.742 | 0.185 | 1.374 | 0.651 | 0.159 | 0.130 |
| 0.568 | 0.081 | 0.344 | 0.517 | 0.337 | 0.720 | 0.126 | 0.027 | 0.167 | 0.159 | 0.445 |
| 0.666 | 0.468 | 0.543 | 0.411 | 0.623 | 1.057 | 0.433 | 0.114 | 1.226 | 1.009 | 0.310 |
| 0.568 | 0.457 | 0.918 | 1.254 | 0.916 | 1.180 | 0.748 | 0.016 | 1.508 | 0.797 | 0.321 |
| 1.195 | 0.239 | 0.001 | 0.495 | 0.383 | 1.301 | 0.744 | 0.623 | 0.830 | 0.372 | 0.452 |
| 0.191 | 0.527 | 1.135 | 1.676 | 1.290 | -1.650 | 0.371 | -0.840 | 1.880 | 0.372 | 0.952 |
| 0.962 | 0.388 | 0.593 | 0.664 | 0.383 | 0.301 | 0.220 | 0.493 | 0.522 | 0.159 | 0.276 |
| 0.927 | 0.150 | 0.494 | 0.011 | 0.490 | 0.097 | 1.090 | 0.710 | 0.537 | 0.584 | -0.727 |
| 0.893 | 1.857 | 0.346 | 0.264 | 0.303 | 0.312 | 0.217 | 0.450 | 1.654 | 0.797 | -0.452 |
| 0.996 | 3.614 | 1.677 | 1.254 | 0.863 | 1.078 | 1.090 | 0.970 | 1.596 | -0.478 | 0.952 |

2. 求样本的相关矩阵 R .

| | | | | | | | | | | |
|-------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 由 \bar{x} | | | | | | | | | | |
| 1.180 | | | | | | | | | | |
| 1.789 | | | | | | | | | | |
| 1.341 | | | | | | | | | | |
| 1.619 | | | | | | | | | | |
| R | 1.348 | 0.097 | 0.880 | 0.947 | 0.879 | 1.000 | | | | |
| | 0.810 | 0.257 | 0.392 | 0.145 | 0.128 | 0.265 | 1.000 | | | |
| | 0.744 | 0.122 | 0.754 | 0.659 | 0.913 | 0.653 | 0.475 | 1.000 | | |
| | 0.655 | 0.016 | 0.773 | 0.698 | 0.921 | 0.757 | 0.380 | 0.866 | 1.000 | |
| | 0.578 | 0.119 | 0.325 | 0.124 | 0.462 | 0.349 | 0.792 | 0.404 | 0.496 | 1.000 |
| | 0.633 | 0.131 | 0.670 | 0.431 | 0.586 | 0.479 | 0.564 | 0.554 | 0.517 | 0.446 |

3. 求 R 的特征值 λ 及其相应的特征向量。

R 为对称阵, 故其特征值为实数, 且特征向量正交。

表 8.3 R 的特征值及其累计方差贡献率

| | | | | | | | | | | | |
|-----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|--------|
| 特征值 | 6.535 | 1.818 | 1.045 | 0.576 | 0.538 | 0.261 | 0.116 | 0.067 | 0.035 | 0.005 | 0.004 |
| 累计方差贡献率/% | 59.41 | 75.94 | 85.43 | 90.67 | 95.56 | 97.94 | 98.99 | 99.60 | 99.92 | 99.97 | 100.00 |

., 在 1 个因子中, 方差贡献率占 59.41%, 因此我们选择 1 个公共因子就可以了。它们对应的特征向量分别为:

$$e_1 = (0.303, -0.072, 0.358, 0.307, 0.376, 0.328, 0.241, 0.347, 0.349, 0.227, 0.282)'$$

$$e_2 = (0.349, 0.153, 0.185, 0.403, 0.154, 0.315, -0.540, 0.027, 0.105, -0.451, 0.170)'$$

$$e_3 = (0.066, 0.918, -0.180, -0.148, 0.108, -0.058, -0.030, 0.095, 0.225, 0.140, -0.026)'$$

4. 求因子载荷矩阵 A 。

根据式(8.17), 得:

$$A = (\sqrt{\lambda_1} e_1, \sqrt{\lambda_2} e_2, \sqrt{\lambda_3} e_3) = \begin{vmatrix} 0.774\ 6 & -0.470\ 3 & 0.087\ 8 \\ -0.183\ 8 & 0.206\ 8 & 0.938\ 1 \\ 0.914\ 5 & 0.249\ 2 & -0.184\ 4 \\ 0.784\ 7 & 0.542\ 7 & -0.151\ 2 \\ 0.960\ 2 & 0.207\ 6 & 0.110\ 5 \\ 0.839\ 5 & 0.425\ 2 & -0.059\ 6 \\ 0.615\ 7 & -0.727\ 9 & -0.031\ 4 \\ 0.888\ 1 & 0.036\ 5 & 0.096\ 7 \\ 0.893\ 4 & 0.141\ 0 & 0.229\ 9 \\ 0.580\ 2 & -0.608\ 2 & 0.143\ 4 \\ 0.721\ 0 & -0.229\ 0 & -0.027\ 0 \\ 0.828\ 9 & & \\ 0.956\ 5 & & \\ 0.932\ 4 & & \\ 0.933\ 2 & & \\ 0.977\ 3 & & \\ 0.889\ 1 & & \\ 0.910\ 0 & & \\ 0.799\ 4 & & \\ 0.871\ 0 & & \\ & & \\ 0.573\ 0 & & \end{vmatrix}$$

共同度,

$$h = \begin{vmatrix} 0.889\ 1 \\ 0.910\ 0 \\ 0.799\ 4 \\ 0.871\ 0 \\ \vdots \\ 0.573\ 0 \end{vmatrix}$$

各变量均服从正态分布,且两两独立,故可作因子分析。

5. 对因子载荷矩阵 A 作正交旋转。

由式 (10-10) 可知, 当因子载荷矩阵 A 的列向量 F_1, F_2, F_3 与公共因子 F_1, F_2, F_3 的夹角 $\theta_1, \theta_2, \theta_3$ 相等时, 公共因子 F_1, F_2, F_3 的方差贡献率相等, 即 $\cos^2 \theta_1 = \cos^2 \theta_2 = \cos^2 \theta_3$ 。

(1) 第一轮循环过程如下:

①对 A 的第 1, 2 列进行如下旋转, 取:

$$\theta = 0.5970$$

$$\text{则 } P = \begin{pmatrix} 0.8270 & -0.5621 \\ 0.5621 & 0.8270 \end{pmatrix}$$

$$AP = \begin{pmatrix} 0.3763 & 0.8244 & 0.0878 \\ 0.0358 & 0.2743 & 0.9381 \\ 0.8964 & -0.3080 & -0.1844 \\ 0.9541 & 0.0077 & -0.1512 \\ 0.9108 & -0.3680 & 0.1105 \\ 0.9333 & -0.1203 & -0.0596 \\ 0.1000 & -0.9481 & -0.0314 \\ 0.7550 & -0.4690 & 0.0967 \\ 0.8182 & -0.3856 & 0.2299 \\ 0.1380 & -0.8291 & 0.1434 \\ 0.4676 & 0.5947 & 0.0270 \end{pmatrix}$$

②对 A 的第 1, 3 列进行如下旋转, 取:

$$\theta = 0.0109$$

$$\text{则 } P = \begin{pmatrix} 0.9999 & 0.0109 \\ -0.0109 & 0.9999 \end{pmatrix}$$

| | | | |
|------|----------|----------|--------|
| | 0.375 3 | 0.824 4 | 091 9 |
| | -0.046 1 | 0.274 3 | 937 6 |
| | 898 4 | -0.308 0 | .174 6 |
| | 955 7 | 0.007 7 | .140 8 |
| | 909 6 | -0.368 0 | .120 5 |
| AP | 933 9 | -0.120 3 | .049 4 |
| | 100 4 | -0.948 1 | .030 3 |
| | 753 9 | -0.469 0 | 104 9 |
| | 815 6 | -0.385 6 | 238 8 |
| | 136 4 | -0.829 1 | 144 9 |
| | 0.467 8 | -0.594 7 | .021 9 |

3. 对 A 的第 2, 3 列进行如下旋转, 取:

$$= 0.155 6$$

$$\text{则 } P = \begin{pmatrix} 0.987 9 & 0.154 9 \\ -0.154 9 & 0.987 9 \end{pmatrix}$$

| | | | |
|-----------|----------|----------|--------|
| | 0.375 3 | -0.828 7 | .036 9 |
| | -0.046 1 | 0.125 7 | 968 8 |
| | 0.898 4 | 0.277 2 | .220 2 |
| | 0.955 7 | 0.029 5 | 137 9 |
| | 0.909 6 | -0.382 2 | .062 0 |
| AP^{-1} | 0.933 9 | -0.111 2 | .067 4 |
| | 0.100 4 | -0.932 0 | .176 8 |
| | 0.753 9 | -0.479 6 | .031 0 |
| | 0.815 6 | -0.417 9 | .176 2 |
| | 0.136 4 | 0.841 6 | .014 7 |
| | 0.467 8 | 0.584 1 | .113 7 |

4. 求 $\varphi_1 = \frac{1}{\sqrt{2}}(A_{11} + A_{21}) = \frac{1}{\sqrt{2}}(0.375 3 + 0.824 4) = 0.600 0$ (各列平方和)

$$\varphi_1 = 0.366 1$$

| | | | |
|---------|----------|----------|-----------|
| | 0.375 3 | -0.828 7 | -0.036 9 |
| | -0.046 1 | 0.125 7 | 0.968 8 |
| | 898 4 | -0.277 2 | -0.220 2 |
| | 955 7 | 0.029 5 | -0.137 9 |
| | 909 6 | 0.382 2 | 0.062 0 |
| A_1^T | 933 9 | -0.111 2 | -0.067 4 |
| | 100 4 | -0.932 0 | -0.0176 8 |
| | 753 9 | 0.479 6 | 0.031 0 |
| | 815 6 | 0.417 9 | 0.176 2 |
| | 136 4 | -0.841 6 | 0.014 7 |
| | 0.467 8 | 0.584 1 | 0.113 7 |

由表 4-1 各元素 x_{ij} 有 $\sum_{j=1}^3 x_{ij}^2 = 10.000 0$, 故 $\sigma_i^2 = 10.000 0/3 = 3.333 3$, 于是 A_1^T 是正交阵, 故 A_1 是正交阵, 故 A_1 关于线性信源各列的方差为:

| | | | |
|---|-----------------------|----------|----------|
| | $\sigma_2 = 0.366\ 3$ | | |
| | 0.367 8 | -0.832 1 | -0.033 8 |
| | -0.056 8 | 0.124 5 | 0.968 1 |
| | 0.898 3 | -0.285 4 | -0.209 7 |
| | 0.957 5 | 0.020 6 | -0.126 1 |
| | 0.905 1 | -0.391 0 | 0.072 5 |
| 1 | 0.933 6 | -0.119 9 | -0.056 2 |
| | 0.093 7 | -0.932 6 | -0.177 3 |
| | 0.748 9 | -0.486 8 | 0.039 4 |
| | 0.809 4 | 0.426 0 | 0.185 4 |
| | 0.128 2 | 0.842 9 | 0.014 8 |
| | 0.463 6 | -0.588 4 | -0.109 1 |

由表 4-1 各元素 x_{ij} 有 $\sum_{j=1}^3 x_{ij}^2 = 10.000 0$, 故 $\sigma_i^2 = 10.000 0/3 = 3.333 3$, 于是 A_2^T 是正交阵, 故 A_2 是正交阵, 故 A_2 关于线性信源各列的方差为:

$$\sigma_2^2 = 0.366 3$$

| | | | |
|------------|----------|----------|---------|
| | 0.366 8 | 0.832 6 | 0.033 8 |
| | -0.056 9 | 0.124 4 | 0.968 1 |
| | 0.898 0 | -0.286 5 | 0.209 1 |
| | 0.957 6 | 0.019 4 | 0.125 8 |
| | 0.904 5 | 0.392 2 | 0.072 7 |
| A_3^{-1} | 0.933 5 | -0.121 1 | 0.055 9 |
| | 0.092 6 | -0.932 7 | 0.177 4 |
| | 0.718 2 | 0.487 8 | 0.039 5 |
| | 0.808 8 | 0.427 0 | 0.185 6 |
| | 0.127 2 | -0.843 0 | 0.014 8 |
| | 0.462 9 | -0.588 9 | 0.109 0 |

由上表可知, 各污染因子在因子1上的得分均较高(因子1的方差为1.086 6), 说明各污染因子在因子1上的得分均较高, 因此因子1为第一主成分, 即污染因子1。

由上表可知, 各污染因子在因子2上的得分均较低(因子2的方差为0.033 8), 说明各污染因子在因子2上的得分均较低, 因此因子2为第二主成分, 即污染因子2。

6. 求因子得分。

(1) 求特殊向量方差 ψ 。

| | |
|---------|---------|
| 0.171 1 | 0.000 0 |
| 0.043 1 | 0.000 0 |
| 0.167 6 | 0.000 0 |
| 0.066 8 | 0.000 0 |
| 0.022 7 | 0.000 0 |
| 0.110 9 | 0.000 0 |
| 0.090 1 | 0.000 0 |
| 0.200 6 | 0.000 0 |
| 0.129 1 | 0.000 0 |
| 0.272 9 | 0.000 0 |
| 0.427 0 | 0.000 0 |

(2) 运用式(8.30), 得到因子得分 F , 结果见表 8.4。

因子得分表

[illegible]

表 8.5 某地区地表水水质监测统计结果 单位: mg/l (pH 除外)

| 样本
编号 | 项目名称 | | | | | | | | | |
|----------|-------|-------|-------|-------|---------|-------|-------|---------|-------|----------|
| | X_1 | X_2 | X_3 | X_4 | X_5 | X_6 | X_7 | X_8 | X_9 | X_{10} |
| 1 | 7.29 | 4.01 | 15.7 | 0.104 | 0.019 5 | 5.62 | 0.097 | 0.008 1 | 0.18 | 0.05 |
| 2 | 7.16 | 6.83 | 19.9 | 0.219 | 0.071 9 | 2.34 | 0.118 | 0.008 1 | 0.08 | 0.06 |
| 3 | 7.19 | 6.00 | 15.5 | 0.063 | 0.016 5 | 5.72 | 0.103 | 0.006 8 | 0.15 | 0.08 |
| 4 | 7.23 | 5.83 | 18.6 | 0.147 | 0.055 6 | 3.89 | 0.111 | 0.007 4 | 0.44 | 0.06 |
| 5 | 7.19 | 4.40 | 16.3 | 0.066 | 0.028 5 | 3.49 | 0.122 | 0.008 1 | 0.24 | 0.10 |
| 6 | 7.01 | 9.38 | 24.3 | 0.236 | 0.032 4 | 1.29 | 0.154 | 0.006 8 | 0.20 | 0.12 |

解 对表 8.5 数据作标准化处理, 所得批标准化后的数据用小写字母 x 表示。

| | | | | | | | | | |
|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 1.189 | -1.070 | -0.793 | -0.468 | -0.820 | 1.076 | -1.019 | 0.858 | -0.285 | -1.044 |
| -0.195 | 0.391 | 0.448 | 1.063 | 1.580 | -0.787 | 0.025 | 0.858 | -1.101 | 0.675 |
| 0.124 | -0.039 | -0.852 | -1.014 | -0.957 | 1.133 | -0.720 | -1.170 | -0.530 | 0.061 |
| 0.124 | -0.868 | -0.616 | -0.974 | -0.408 | -0.133 | 0.224 | 0.858 | 0.204 | 0.798 |
| -1.793 | 1.712 | 1.749 | 1.289 | -0.229 | -1.383 | 1.814 | -1.170 | -0.122 | 1.535 |

2. 求样本的相关矩阵 R 。

| | | | | | | | | | |
|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 1.000 | -0.920 | -0.863 | -0.653 | -0.117 | 0.818 | -0.941 | 0.557 | 0.213 | -0.825 |
| -0.920 | 1.000 | 0.916 | 0.799 | 0.278 | -0.753 | 0.816 | -0.651 | -0.145 | 0.574 |
| -0.863 | 0.916 | 1.000 | 0.915 | 0.416 | -0.901 | 0.889 | -0.378 | -0.007 | 0.525 |
| -0.653 | 0.799 | 0.915 | 1.000 | 0.645 | -0.826 | 0.664 | -0.134 | -0.133 | 0.168 |

R

| | | | | | | | | | |
|-------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 0.818 | -0.753 | -0.901 | -0.826 | -0.552 | 1.000 | 0.900 | 0.081 | 0.054 | -0.563 |
| 0.541 | 0.816 | 0.889 | 0.664 | 0.158 | -0.900 | 1.000 | 0.373 | -0.025 | 0.837 |
| 0.557 | -0.651 | -0.378 | -0.134 | 0.299 | 0.081 | -0.373 | 1.000 | -0.136 | -0.500 |
| 0.213 | -0.145 | 0.007 | -0.133 | 0.095 | 0.054 | -0.025 | -0.136 | 1.000 | -0.051 |
| 0.825 | 0.574 | 0.525 | 0.168 | -0.302 | -0.563 | 0.837 | 0.500 | -0.051 | 1.000 |

3. 求 R 的特征值 λ 及其相应的特征向量。

R 的特征值及其累计方差贡献率, 见表 8.6。

表 8.6

特征值及其累计方差贡献率

| 特征值 | 5.952 4 | 1.979 7 | 1.090 0 | 0.765 6 | 0.212 3 |
|-----------|---------|---------|---------|---------|---------|
| 累计方差贡献率/% | 59.52 | 79.32 | 90.22 | 97.88 | 100.00 |

由表 8.6 可知, 前 3 个特征值之和占全部特征值的 90.22%, 因此, 选取前 3 个特征向量作为主成分, 即主成分 F_1, F_2, F_3 。主成分 F_1 反映了水质污染的综合水平, 主成分 F_2 反映了水质污染的综合水平, 主成分 F_3 反映了水质污染的综合水平。因此, 主成分 F_1, F_2, F_3 可以代表水质污染的综合水平。

它们对应的特征向量:

$$\begin{aligned}
 (e_1, e_2, e_3) = & \begin{pmatrix} 0.394 & -0.163 & -0.116 \\ 0.388 & 0.047 & -0.012 \\ 0.396 & 0.115 & 0.084 \\ 0.335 & 0.352 & 0.026 \\ 0.132 & -0.624 & -0.150 \\ 0.373 & 0.214 & -0.024 \\ 0.389 & 0.093 & -0.005 \\ 0.189 & -0.457 & 0.292 \\ -0.039 & -0.004 & -0.931 \\ 0.283 & 0.425 & 0.054 \end{pmatrix}
 \end{aligned}$$

4. 求因子载荷矩阵 A 。

根据式(8.17):

$$\begin{aligned}
 A = (\sqrt{\lambda_1} e_1, \sqrt{\lambda_2} e_2, \sqrt{\lambda_3} e_3) = & \begin{pmatrix} 0.961 & 0.230 & 0.122 \\ 0.946 & 0.067 & 0.013 \\ 0.965 & 0.161 & 0.088 \\ 0.817 & -0.496 & 0.027 \\ 0.323 & -0.879 & -0.156 \\ 0.910 & 0.301 & -0.025 \\ 0.949 & 0.131 & -0.006 \\ 0.462 & 0.642 & 0.305 \\ 0.095 & -0.005 & -0.972 \\ 0.692 & 0.599 & 0.057 \end{pmatrix}
 \end{aligned}$$

【思考题8】

1. 试述因子分析的基本思想
2. 比较因子分析和主成分分析模型的关系, 说明其异同点。

× 循环经济发展情况进行分类。

- (1) 求样本的相关矩阵 R ;
- (2) 求 R 的特征值及其累计方差贡献率;
- (3) 求因子载荷矩阵、共同度及各个公因子 f_i 对所有变量的贡献;
- (4) 求因子得分, 并进行分析

表 8.6

各省份循环经济发展情况

| 省份 | 单位 GDP 用水量 (m ³ · 万元 ⁻¹) | 单位 GDP 能耗 (吨标煤 · 万元 ⁻¹) | 单位 GDP 二氧化碳排放量 (吨 · 万元 ⁻¹) | “三废”综合利用率 (%) | 环保投资占 GDP 比重 (%) |
|------|---|-------------------------------------|--|---------------|------------------|
| 北京 | 1.29 | 80.78 | 2 610.02 | 0.37 | 1.53 |
| 上海 | 1.07 | 158.52 | 9 042.69 | 0.22 | 0.94 |
| 广东 | 0.96 | 289.79 | 892.29 | 0.38 | 0.70 |
| 山西 | 4.23 | 183.74 | 194.14 | 0.83 | 1.48 |
| 安徽 | 1.47 | 435.72 | 343.45 | 0.67 | 0.86 |
| 湖北 | 1.42 | 384.63 | 339.45 | 1.45 | 0.71 |
| 重庆 | 1.32 | 253.25 | 323.98 | 0.54 | 1.81 |
| 内蒙古 | 2.43 | 632.36 | 23.68 | 0.61 | 1.63 |
| 甘肃 | 2.70 | 781.31 | 38.58 | 0.76 | 1.06 |
| 全国平均 | 1.43 | 405.32 | 143.98 | 0.64 | 1.40 |

数据来源: 中国统计年鉴、能源统计年鉴和环境统计年鉴。

- (1) 求样本的相关矩阵 R ;
- (2) 求 R 的特征值及其累计方差贡献率;

业区进行分类评价。

- (1) 求样本的相关矩阵 R ;
- (2) 求 R 的特征值及其累计方差贡献率;
- (3) 解释各公因子所代表的意义;
- (4) 求因子得分, 并进行分类评价。

【参考文献】

社, 1988.

表 9.1 人工神经网络发展的几个阶段

| 阶段 | 时间/年 | 代表人物/会议 | 主要贡献 |
|------------------|-------|-------------------------|--|
| 早期阶段 | 1943 | W. McCulloch 和 W. Pitts | 提出了 MP 模型 |
| | 1948 | Wiener | 提出了伺服机反馈自稳定系统的概念 |
| | 1949 | 心理学家 D.O. Hebb | 在 <i>The Organization of Behavior</i> 书中提出了著名的 Hebb 学习规则 |
| | 1957 | F. Rosenblatt | 提出了感知器(perception)模型 |
| | | W. S. McCulloch | 第一次把神经网络研究从纯理论的研究付诸工程应用,掀起了神经网络研究的第一次高潮 |
| | | McCulloch 和 Pitts | 胞有限自动机的理论模型 |
| 过
渡
阶
段 | 1972 | T. Kohonen | 提出了联想记忆理论 |
| | ~1977 | | 任意可视模式的联想问题上 |
| | | 家 J. Hopfield | 神经网络研究高潮的又一次到来 |
| | | | 神经网络研究高潮的又一次到来 |
| | | | 神经网络研究高潮的又一次到来 |
| | | | 神经网络研究高潮的又一次到来 |
| 现代阶段 | | 一次国际神经网络会议 | 研究迅速发展起来 |
| | | | 神经网络仿真开发环境 |

ANN 具有多种类型,它仅用少数简单元件的一种或几种传输函数,产生算法三大要素。ANN 具有以下显著的特点:

入、输出、传递函数等, 模型中利用非线性或非线性转移特性 (如饱和、延迟) 等, 输入与输出、反馈等, 模型尚未定义, 它通过(多个)输入(输出)和模型中的非线性函数(非线性函数变换)实现的。图 9-1



图 9-1 人工神经元基本结构图

$$y = f\left(\sum_{i=1}^n w_i x_i - \theta\right)$$

$f(\cdot)$ 一般取下面二种函数:

(1) 线性传递函数 (图 9-2)

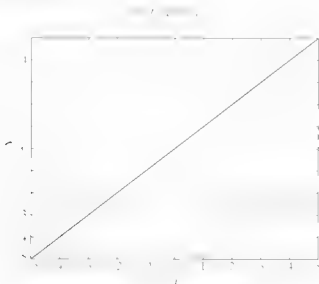


图 9-2 线性传递函数图

(2) 双曲正切 S 型传递函数 (图 9-3)

$$y=f(a)=\frac{1-\exp(-2a)}{1+\exp(-2a)}$$

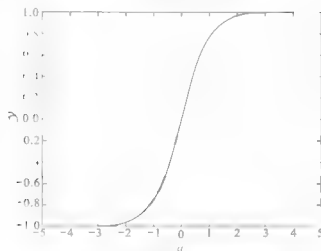


图 9-3 双曲正切 S 型传递函数图

(3) 对数 S 型传递函数 (图 9-4)

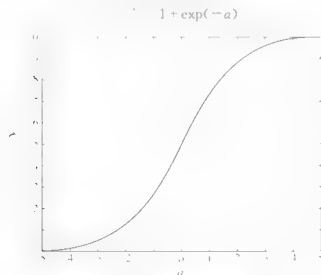


图 9-4 对数 S 型传递函数图

9.3 BP 神经网络

神经网络 (Neural Network) 是模仿人脑神经网络而设计的一种非线性映射模型, 它由大量的简单单元 (神经元) 组成, 通过大量的连接 (突触) 实现信息传递。在数学上, 神经网络可以看作是一种非线性映射模型, 它可以将输入空间的低维数据映射到高维空间, 从而实现非线性映射。神经网络在 ANN 中占有重要地位, 它是 ANN 中最重要、应用最广泛的模型之一。神经网络具有强大的非线性映射能力, 可以用于解决各种非线性问题, 如模式识别、函数逼近、分类、回归等。

9.3.1 BP 神经网络原理

BP 神经网络 (Backpropagation Neural Network) 是一种典型的神经网络, 其结构如图 9-5 所示。



图 9-5 BP 网络的拓扑结构

神经网络可以看作是一种非线性映射模型, 它可以将输入空间的低维数据映射到高维空间, 从而实现非线性映射。

定理 1 存在一个神经网络, 可以任意逼近任意一个连续函数 (Hornik et al., 1989; Cybenko, 1989; Universal Approximation Property, 1995)。

定理 2 存在一个神经网络, 可以任意逼近任意一个连续函数 (Hornik et al., 1989; Cybenko, 1989; Universal Approximation Property, 1995)。

定理 3 存在一个神经网络, 可以任意逼近任意一个连续函数 (Hornik et al., 1989; Cybenko, 1989; Universal Approximation Property, 1995)。

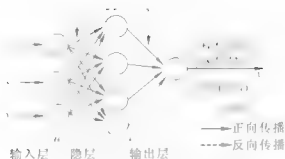


图 9-6 BP 算法原理示意图

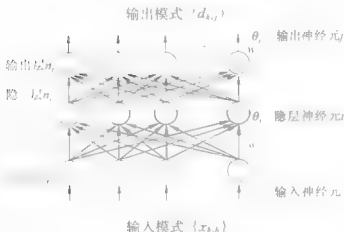


图 9-7 BP 神经网络的拓扑结构

中 W_{ij} 为连接权, 计算 h_j 值, $j=1, 2, \dots, n_h$ 。

步骤 1: 归一化。对输入层各节点权值 W_{ij} 归一化处理, 并对节点输入、输出权值作归一化处理。设输入层节点数为 n_i , $i=1, 2, \dots, n_i$; 隐层节点数为 n_h , $h=1, 2, \dots, n_h$; 输出层节点数为 n_o , $o=1, 2, \dots, n_o$ 。各节点权值 W_{ij} 和阈值 $\{\theta_j\}$, $\{\theta_j\}$ 赋予 $(-0.1, 0.1)$ 区间上的随机值。

步骤 2: 在输入层各节点 $x_{i,k}$ 处, 按前向网络 $i=1, 2, \dots, n_i$, $k=1, 2, \dots, n_k$ 。

步骤 3: 计算隐层各节点输入 h_j , $j=1, 2, \dots, n_h$ 。

$$x_i = \sum_{k=1}^n w_{ik} x_{k,h} + \theta_i$$

$$y_i = f_1(x_i)$$

其中, $f_1(\cdot)$ 是隐层各节点神经元的激活函数。

步骤4: 计算隐层各节点神经元的净输入 x_i ($i=1, 2, \dots, n$)。

$$x_j = \sum_{i=1}^n w_{ij} y_i + \theta_j$$

$$y_j = f_2(x_j)$$

其中, $f_2(\cdot)$ 是输出层各节点神经元的激活函数。

计算第 k 个单样本点的误差:

$$E_k = \sum_j (y_j - d_{k,j})^2 / 2$$

图 9-8 为 BP 算法中 E_k 与其他变量之间的函数关系示意图。

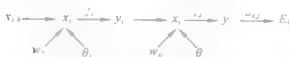


图 9-8 BP 算法中 E_k 与其他变量之间的函数关系示意图

计算 E_k 对 x_j 的偏导数 $\frac{\partial E_k}{\partial x_j}$ 和 E_k 对 y_j 的偏导数 $\frac{\partial E_k}{\partial y_j}$ 。

步骤5: 计算 E_k 对 w_{ij} 的偏导数 $\frac{\partial E_k}{\partial w_{ij}}$ 和 E_k 对 θ_j 的偏导数 $\frac{\partial E_k}{\partial \theta_j}$ 。

$$\begin{aligned} \text{grad}_{w_{ij}}(E_k) &= \frac{\partial E_k}{\partial w_{ij}} = \frac{\partial E_k}{\partial x_j} \cdot \frac{\partial x_j}{\partial w_{ij}} \\ &= [(y_j - d_{k,j}) \cdot f'_2(x_j)] \cdot y_i \\ &= \delta_{k,j} \cdot y_i \end{aligned}$$

$$\Delta w_{ij} = -\eta \cdot \frac{\partial E_k}{\partial w_{ij}}$$

$$\begin{aligned} \text{grad}_{\theta_j}(E_k) &= \frac{\partial E_k}{\partial \theta_j} = \frac{\partial E_k}{\partial x_j} \cdot \frac{\partial x_j}{\partial \theta_j} \\ &= [(y_j - d_{k,j}) \cdot f'_2(x_j)] \cdot 1 \\ &= \delta_{k,j} \end{aligned}$$

$$\Delta \theta_j = -\eta \cdot \frac{\partial E_k}{\partial \theta_j}$$

步骤6: 计算隐层各节点神经元的修正量 Δw_{ik} 和 $\Delta \theta_i$ 。

$$\begin{aligned}
 \text{grad}_{w_h}(E_k) &= \frac{\partial E_k}{\partial w_h} = \frac{\partial E_k}{\partial x_i} \cdot \frac{\partial x_i}{\partial w_h} \\
 &= \left(\sum_{j=1}^{n_j} \frac{\partial E_k}{\partial x_j} \cdot \frac{\partial x_j}{\partial y_i} \cdot \frac{\partial y_i}{\partial x_i} \right) \cdot \frac{\partial x_i}{\partial w_h} \\
 &= \left[\sum_{j=1}^{n_j} \delta_{k,j} \cdot w_{ij} \cdot f'_{ij}(x_i) \right] \cdot x_{k,h}
 \end{aligned}$$

$$\begin{aligned}
 \Delta w_{h,i} &= -\eta \cdot \frac{\partial E_k}{\partial w_{h,i}} \\
 \text{grad}_{\theta_i}(E_k) &= \frac{\partial E_k}{\partial \theta_i} = \frac{\partial E_k}{\partial x_i} \cdot \frac{\partial x_i}{\partial \theta_i} \\
 &= \sum_{j=1}^{n_j} \hat{\theta}_{k,j} \cdot w_{ij} \cdot f'_{ij}(x_i) \\
 \hat{w}_{ij} &= \hat{\theta}_{k,j} \\
 \Delta \theta_i &= -\eta \cdot \frac{\partial E_k}{\partial \theta_i}
 \end{aligned}$$

步骤7: 修正各连接的权值和阈值。

$$\begin{aligned}
 w_{ij}^{t+1} &= w_{ij}^t + \Delta w_{ij} \\
 \theta_i^{t+1} &= \theta_i^t + \Delta \theta_i \\
 w_{h,i}^{t+1} &= w_{h,i}^t + \Delta w_{h,i} \\
 \theta_i^{t+1} &= \theta_i^t + \Delta \theta_i
 \end{aligned}$$

由于学习速率 η 对收敛速度影响较大, 若 η 较大, 则算法收敛快, 但不稳定; 若 η 较小, 则算法收敛慢, 但收敛较稳定。

步骤8: 由式(10)计算各模式对 (x, d) 提供给网络, 转步骤9, 直至全部 n_s 个模式对训练完毕, 转步骤9。

步骤9: 计算误差函数 E , 计算可各个训练误差函数



图 9-10 解异或问题 BP 网络的拓扑结构

图 9-11 为 BP 算法(2-2-1)的训练误差图。由图可见, BP 算法(2-2-1)的训练误差在 500 次迭代后, 误差平方和已降至 10 以下。

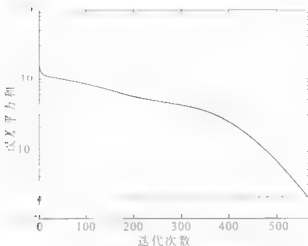


图 9-11 BP 算法(2-2-1)的训练误差图

BP 算法各层训练结果如下:

(1) 隐层训练结果

$$w_{h1} = 7.743\ 7 \quad -6.692\ 5$$

$$2.759\ 5 \quad 2.086\ 1$$

$$\theta_1 = -1.691\ 1, 7.452\ 5$$

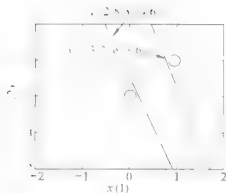


图 9-12 BP 算法(2.2.1)的隐层训练结果图

(2) 输出层训练结果

$$w_{0j} = 1.679\ 1$$

$$1.717\ 2$$

$$\theta_j = -1.950\ 5$$

BP 计算值:

$$y = 0.027\ 1, 1.008\ 5, 0.894\ 7, 0.088\ 6$$

目标值:

$$T = 0, 1, 1, 0$$

BP 网络训练结果, 见图 9-13。

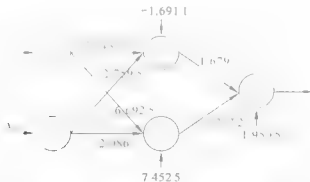


图 9-13 BP 算法(2.2.1)的训练结果图

图 9-13 中, $x(1)$ 为输入, $x(2)$ 为输出, $x(3)$ 为训练结果, $x(4)$ 为目标值, $x(5)$ 为误差。

1. 隐层节点对网络训练的影响

图 9-14 为隐层节点数与训练次数、误差的关系图。由图 9-14 可以看出, 隐层节点数与训练次数、误差的关系见图 9-14。

表 9.2 隐层节点对网络训练的影响

| 隐层节点数 | 迭代次数 | 误差 |
|-------|------|---------|
| 2 | 568 | 0.019 7 |
| 3 | 85 | 0.019 5 |
| 4 | 275 | 0.019 1 |
| 5 | 141 | 0.019 7 |
| 6 | 181 | 0.020 0 |
| 7 | 17 | 0.019 6 |
| 8 | 42 | 0.019 6 |
| 9 | 37 | 0.019 2 |
| 10 | 14 | 0.016 8 |
| 15 | 55 | 0.016 5 |
| 20 | 30 | 0.020 0 |
| 25 | 28 | 0.020 0 |
| 30 | 59 | 0.020 0 |

由图 9-14 可以看出, 当隐层节点数一定时, 训练次数会减少, 当增加到一定数目后, 训练次数又会增加。

2. 学习速率对网络训练的影响

学习速率对网络训练的影响如图 9-15 所示。由图 9-15 可以看出, 学习速率对网络训练的影响。

由图 9-15 可以看出, 当学习速率一定时, 训练次数会减少, 当增加到一定数目后, 训练次数又会增加。因此, 学习速率的选择 0.1 是比较合适的。

综上所述, 可以得到以下结论:

1. 当隐层节点数一定时, 训练次数会减少, 当增加到一定数目后, 训练次数又会增加。

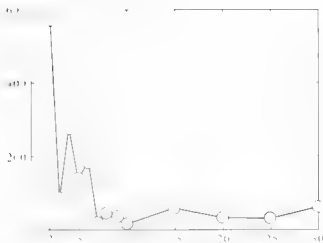


图 9-14 BP 算法的迭代次数与隐层节点的关系图

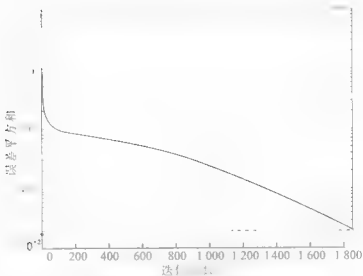


图 9-15 学习速率为 0.01 时网络训练的结果图

BP 神经网络中每个受训节点的输出是其所接收的输入神经元的函数, 因此, 对每个节点的输出函数, 必须事先给定, 而输出函数由设计者来决定。

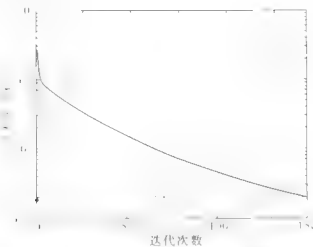


图 9-16 学习速率为 0.1 时网络训练的结果图

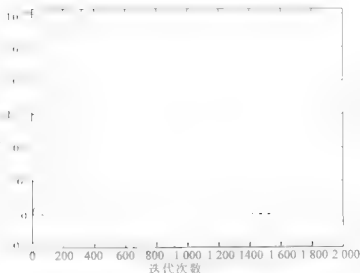


图 9-17 学习速率为 1 时网络训练的结果图

目前神经网络的学习速度较慢,网络设计有待进一步完善,最后,对于采用附加动量法等改进方法来训练网络。

表 9.4 训练样本模拟结果与实测值的比较

| 指标 | 数 据 组 | | | | | | | |
|----------|---------|---------|----------|---------|----------|---------|----------|----------|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 实测值 | 8.4 | 8.5 | 8.3 | 8.4 | 7.0 | 7.4 | 7.5 | 7.3 |
| IX) 预测值 | 8.393 5 | 8.498 5 | 8.300 6 | 8.398 2 | 7.001 7 | 7.398 3 | 7.501 4 | 7.300 4 |
| 绝对误差 | 0.006 5 | 0.001 5 | -0.000 6 | 0.001 8 | 0.001 7 | 0.001 7 | -0.001 4 | -0.000 4 |
| 实测值 | 1.1 | 1.2 | 1.5 | 1.2 | 0.6 | 0.8 | 0.9 | 1.0 |
| IXD) 预测值 | 1.100 1 | 1.200 0 | 1.499 7 | 1.199 3 | 0.600 2 | 0.800 3 | 0.900 4 | 0.999 7 |
| 绝对误差 | 0.000 1 | 0.000 0 | 0.000 3 | 0.000 7 | -0.000 2 | 0.000 3 | 0.000 4 | 0.000 3 |

表 9.6 检测样本预测结果与实测值的比较

| 指标 | 实测值 | 预测值 | 绝对误差 |
|------|-----|---------|----------|
| IX) | 8.5 | 8.307 8 | -0.192 2 |
| | | 7.015 3 | -0.184 7 |
| IXD) | 1.5 | 1.222 3 | 0.277 7 |
| | | 0.9 | 0.117 9 |

9.4 RBF 神经网络

图 9-1 所示为 RBF 神经网络模型。图中, 输入节点与隐含节点在训练过程中, 隐含节点与输出节点之间存在权值, 隐含节点与输出节点之间存在偏置函数, 隐含节点与输出节点之间存在权值, 隐含节点与输出节点之间存在偏置函数。RBF 神经网络无论在学习能力、分类能力、泛化能力、鲁棒性、稳定性和对异常值的鲁棒性, 本节给出 Matlab 环境下的 RBF 网络模型。

9.4.1 RBF 神经网络原理

RBF 网络为两层网络, 第一层为隐含的径向基层, 第二层为输出线性层, 其网络结构如图 9-18 所示。

其中, σ_i^2 为常数, $\sigma_i^2 = \sigma_i^2(x, y)$ 是 σ_i^2 的泛函, σ_i^2 为常数。

$$R_i(x) = \exp\left(-\frac{\|x - c_i\|^2}{2\sigma_i^2}\right) \quad (i=1, 2, \dots, p)$$

其中, x 为输入, c_i 为第 i 个感知单元的质心, σ_i^2 为第 i 个感知单元的方差, p 是感知单元的个数, $\|x - c_i\|$ 是向量 $x - c_i$ 的范数。



图 9-18 RBF 网络结构

从式(9-18)可知, 输入 x 与输出 y 的关系可表示为: 线性映射、非线性映射从 $R_i(x) \rightarrow y_i$ 的线性映射, 即:

$$\sum_{i=1}^q k_i x_i \quad (i=1, 2, \dots, q)$$

其中, q 是输出节点数。

从式(9-19)可知, k_i 可看作 y_i 的权值, 与 x_i 相乘。

9.4.2 RBF 神经网络模型

RBF 网络具有非线性、局部敏感、收敛快、可训练、可推广等特点。为了能够利用 RBF 神经网络对非线性问题, 本章直接采用与神经网络原理相同的 RBF 神经网络模型。

神经网络模型由输入层、隐含层和输出层组成。输入层和隐含层为无监督学习层, 输出层为有监督学习层。神经网络模型采用 AR 模型。多输入、多输出、无监督、有监督、非线性、局部敏感、收敛快、可训练、可推广等特点。

神经网络模型由输入层、隐含层和输出层组成。输入层和隐含层为无监督学习层,

例 9.4 本厂于 1997 年 12 月 1 日购入一台设备, 其原值为 100 000 元, 预计使用年限为 5 年, 预计净残值为 5 000 元。按直线法计提折旧。至 1999 年 12 月 31 日, 该设备已使用 2 年, 累计折旧额为 38 000 元。现因该设备陈旧, 决定提前报废。假设该设备已计提减值准备 10 000 元。请编制该设备报废时的会计分录。

表 9-7 某海羊冰情等级序列观测值和各模型的拟合结果与预测结果 单位 冰级

注: * 表示 TAR 的绝对误差, ** 表示 RBF 方法的绝对误差。

表 9-8 某海洋冰情等级序列自相关系数及其上、下限值 (置信水平 70%)

| k | 1 | 2 | 3 | 4 | 5 | 6 |
|----------|-------|--------|-------|-------|--------|-------|
| $R_1(k)$ | 0.238 | -0.244 | 0.249 | 0.256 | 0.262 | 0.269 |
| $R(k)$ | 0.251 | 0.105 | 0.217 | 0.278 | -0.100 | 0.110 |
| $R(k)$ | 0.162 | 0.164 | 0.166 | 0.169 | 0.171 | 0.174 |

算过程如下:

1. 用式(9.2)建立 $n-m$ 组训练样本的输入、输出向量。

这里, $n=27$, $m=4$ 。

2. 设计 RBF 网络。令 $g=0.000\ 01$, $s=1$

$$\text{net}=\text{newrb}(x, \bar{y}, g, s)$$

3. 由 $t=\text{sim}(\text{net}, x)$, 可得网络的训练结果。

值, 用“○”表示。从图 9-19 可以看出, 计算误差为 0。

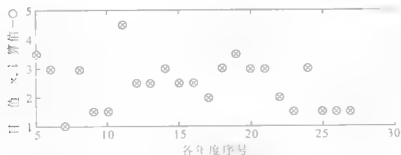


图 9-19 网络的训练结果

由 $a=(3.0, 1.5, 1.5, 1.5)$ 和 $b=\text{sim}(\text{net}, a)$, 可得到 1993~1994 年度冰情等级的预测结果 $b=1.756\ 6$, 详见表 9.7。

在表 10-1-1 中, 第 1 列为输入变量, 其余列为历史预报。从表 10-1-1 可知, 在 1997 年 1 月 1 日以前, 输入变量为 1996 年 12 月 31 日的数据, 预报不包括当天数据, 而试验结果 \hat{y}_t 为 1997 年 1 月 1 日的数据。ARIMA 模型和灰色模型、RBF 神经网络模型均利用 1996 年 12 月 31 日的数据进行建模, 因此, 在 1997 年 1 月 1 日进行预报时, 输入变量 x_t 为空值, 而试验结果 \hat{y}_t 为 1997 年 1 月 1 日的数据, 其他输入变量和输出变量都是令人满意的。

【思考题 9】

试述 BP 神经网络原理

BP 神经网络采用的激发函数为什么必须是连续可导的?

给出 BP 算法的基本步骤及计算框图

试述 BP 算法的优缺点。

试述 RBF 神经网络原理及计算框图

试述 RBF 神经网络的优缺点。

试比较 BP 与 RBF 神经网络的性能。

试用 BP 神经网络解决一个实际环境问题

试用 RBF 神经网络解决一个实际环境问题

【参考文献】

1. 李德明, 王德明. 灰色系统理论及其在环境科学中的应用[M]. 北京: 中国环境出版社, 2005.
2. 李德明, 王德明. 灰色系统理论及其在环境科学中的应用[M]. 北京: 中国环境出版社, 2003.
3. 李德明, 王德明. 灰色系统理论及其在环境科学中的应用[M]. 北京: 中国环境出版社, 2003.
4. 李德明, 王德明. 灰色系统理论及其在环境科学中的应用[M]. 北京: 中国环境出版社, 2003.
5. 李德明, 王德明. 灰色系统理论及其在环境科学中的应用[M]. 北京: 中国环境出版社, 2000.
6. 李德明, 王德明. 灰色系统理论及其在环境科学中的应用[M]. 北京: 中国环境出版社, 2000.

- 李士勇, 1999, 21(2): 140-143.
- 李士勇, 1999, 21(2): 140-143.
- (4): 70-75.
- 李士勇, 1999, 14(4): 1-6.
- 李士勇, 1999, 18(3): 1-6.
- 李士勇, 2007, 21(1): 38-44.

环境统计分析

环境统计工作的发展,在历史上曾一度处于停滞、徘徊状态,更有多数分支,环境信息系统就是其中的典型。

3. 社会来源

环境统计工作的发展,在历史上曾一度处于停滞、徘徊状态,更有多数分支,环境信息系统就是其中的典型。环境统计工作的发展,在历史上曾一度处于停滞、徘徊状态,更有多数分支,环境信息系统就是其中的典型。环境统计工作的发展,在历史上曾一度处于停滞、徘徊状态,更有多数分支,环境信息系统就是其中的典型。

环境统计工作的发展,在历史上曾一度处于停滞、徘徊状态,更有多数分支,环境信息系统就是其中的典型。环境统计工作的发展,在历史上曾一度处于停滞、徘徊状态,更有多数分支,环境信息系统就是其中的典型。环境统计工作的发展,在历史上曾一度处于停滞、徘徊状态,更有多数分支,环境信息系统就是其中的典型。

10.2 环境空间统计分析

环境空间统计分析是环境统计的重要组成部分,它主要研究环境要素在空间上的分布规律、相互关系及变化过程。常用的方法包括地理信息系统(GIS)、遥感技术、空间插值等。

环境空间统计分析的主要内容包括:空间数据的采集、处理、分析和解释。常用的方法包括:点模式分析、面模式分析、网络模式分析等。此外,还有空间回归分析、空间自相关分析等方法。

环境空间统计分析的理论基础是地理学、统计学和计算机科学。常用的软件包括ArcGIS、MapInfo、SPSS等。在实际应用中,环境空间统计分析可以帮助我们了解环境问题的空间分布规律,为环境管理和决策提供科学依据。

4. 准二阶平稳假设和准本征假设

如果随机函数 $Z(x)$ 的均值 $E[Z(x)] = \mu$ 在域 X 上为常数, 且 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的, 则称 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的。如果 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的, 且 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的, 则称 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的。

如果 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的, 且 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的, 则称 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的。如果 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的, 且 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的, 则称 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的。如果 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的, 且 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的, 则称 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的。如果 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的, 且 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的, 则称 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的。

如果 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的, 且 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的, 则称 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的。如果 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的, 且 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的, 则称 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的。

10.2.3.2 变差函数

如果 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的, 且 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的, 则称 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的。如果 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的, 且 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的, 则称 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的。如果 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的, 且 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的, 则称 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的。

$$\gamma(x, h) = \frac{1}{2} D[Z(x) - Z(x+h)] \quad (10.16)$$

其中, D 为方差, $D[Z(x) - Z(x+h)] = E[Z(x) - Z(x+h)]^2$ 。

$$\begin{aligned} \gamma(x, h) &= \frac{1}{2} D[Z(x) - Z(x+h)] \\ &= \frac{1}{2} E[Z(x) - Z(x+h)]^2 - \frac{1}{2} \{E[Z(x)] - E[Z(x+h)]\}^2 \end{aligned} \quad (10.17)$$

如果 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的, 且 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的, 则称 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的。如果 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的, 且 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的, 则称 $Z(x)$ 在域 X 上为二阶平稳(或准本征)的。

$$E[Z(x+h)] = E[Z(x)], \quad \forall h \quad (10.18)$$

因此, 式(10.17)就可以简化为:

$$\gamma(x, h) = \frac{1}{2} E[Z(x) - Z(x+h)]^2 \quad (10.19)$$

(6) 对数模型。其一般公式为:

$$\gamma(h) = A \lg h \quad (10.27)$$

对数模型不能描述点支撑上的区域化变量的结构。

(7) 线性有基台值模型。其一般公式为:

$$\gamma(h) = \begin{cases} c_0 & (h=0) \\ Ah & (0 < h \leq a) \\ c_0 + c & (h > a) \end{cases} \quad (10.28)$$

该模型的变程为 a , 基台值为 $c_0 + c$ 。

(8) 线性无基台值模型。其一般公式为:

$$\gamma(h) = \begin{cases} c_0 & (h=0) \\ Ah & (h > 0) \end{cases} \quad (10.29)$$

该模型没有基台值, 也没有变程。

10.1 所示。

变差函数是空间统计学的主要工具, 有了变差

空间污染物的空间分布进行了研究。

4. 变差函数的参数最优估计

变差函数的理论模型主要是曲线模型, 将曲线

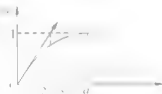


图 10-1 球状模型曲线

表 10.1 常用变差函数理论模型的线性变换

| 变差函数理论模型 | | 变换 | 变换后的线性模型 |
|-------------|----------------------------------|----------------------------|----------|
| 球状模型 | | $\gamma(h) = y, c_0 = b_0$ | |
| 0 | $(h=0)$ | $h = \frac{3c}{a}$ | |
| $\gamma(h)$ | $\frac{3h}{a} - \frac{h^3}{a^3}$ | | |
| $c_0 + c$ | $(h > a)$ | | |

续表

| 变差函数理论模型 | 变换 | 变换后的线性模型 |
|--|---|-------------------|
| 指数模型 | $\gamma(h) = y, e^{\frac{h}{a}} = x$ | |
| $\gamma(h) = \begin{cases} 0 & (h=0) \\ c_0 + c(1 - e^{-\frac{h}{a}}) & (h>0) \end{cases}$ | $\begin{cases} c_0 + c - b_0 \\ -c - b_1 \end{cases}$ | $y = b_0 + b_1 x$ |
| 高斯模型 | $\gamma(h) = y, e^{-\frac{h^2}{a^2}} = x$ | |
| $\gamma(h) = \begin{cases} 0 & (h=0) \\ c_0 + c(1 - e^{-\frac{h^2}{a^2}}) & (h>0) \end{cases}$ | $\begin{cases} c_0 + c = b_0 \\ -c - b_1 \end{cases}$ | $y = b_0 + b_1 x$ |

设二元线性回归模型为:

$$y = b_0 + b_1 x_1 + b_2 x_2 \quad (10.30)$$

其中, y 为因变量, x_1, x_2 为自变量, b_0, b_1, b_2 为待估计的参数。

$$b_0 = y - b_1 x_1 - b_2 x_2$$

$$\begin{bmatrix} 1 & -x_1 & -x_2 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

中,

$$\begin{aligned} y &= \sum_{i=1}^n N(h_i) y_i / \sum_{i=1}^n N(h_i) \\ \bar{x}_1 &= \sum_{i=1}^n N(h_i) x_{1i} / \sum_{i=1}^n N(h_i) \\ \bar{x}_2 &= \sum_{i=1}^n N(h_i) x_{2i} / \sum_{i=1}^n N(h_i) \\ L_{11} &= \sum_{i=1}^n N(h_i) (x_{1i} - \bar{x}_1)^2 \end{aligned}$$

$$\begin{aligned}
 L_{22} &= \sum_{i=1}^n N(h_i)(x_{2i} - x_2)^2 \\
 L_{12} &= L_{21} = \sum_{i=1}^n N(h_i)(x_{1i} - x_1)(x_{2i} - x_2) \\
 S_{yy} &= \sum_{i=1}^n N(h_i)(y_i - \bar{y})(y_i - \bar{y}) \\
 S_{xx} &= \sum_{i=1}^n N(h_i)(x_i - \bar{x})(x_i - \bar{x}) \\
 S_{xy} &= \sum_{i=1}^n N(h_i)(x_i - \bar{x})(y_i - \bar{y})
 \end{aligned}$$

从而可求得拟合球状模型时的三个参数。

如果数据点分布不均匀, 且数据点个数不多, 则可用最小二乘法求得三个参数。

$$\begin{aligned}
 c_0 &= b_0 \\
 a &= \sqrt{\frac{c}{3b_2}} \\
 c &= \frac{2b_1}{3} \sqrt{\frac{-b_1}{3b_2}}
 \end{aligned} \quad (10.32)$$

这三个参数为最优拟合球状模型时的三个参数。

如果数据点分布不均匀, 且数据点个数不多, 则可用最小二乘法求得三个参数。如果数据点分布不均匀, 且数据点个数不多, 则可用最小二乘法求得三个参数。如果数据点分布不均匀, 且数据点个数不多, 则可用最小二乘法求得三个参数。

如果数据点分布不均匀, 且数据点个数不多, 则可用最小二乘法求得三个参数。如果数据点分布不均匀, 且数据点个数不多, 则可用最小二乘法求得三个参数。如果数据点分布不均匀, 且数据点个数不多, 则可用最小二乘法求得三个参数。

5. 回归模型的检验

回归模型建立后, 需要对模型进行检验, 仅仅进行系数检验是不够的, 还必须对模型的拟合度进行检验, 这称为模型的拟合度检验。

拟合度检验的方法有多种, 这里介绍一种称为平方和检验的方法。

平方和检验的基本思想是: 将总平方和分解为回归平方和和残差平方和, 从而得到回归平方和和残差平方和, 从而得到回归平方和和残差平方和。

$$Q = \sum (y_i - \hat{y}_i)^2 \quad (10.33)$$

1. 在空间上, 数据是随机的, 即数据在空间上的分布是随机的, 没有明显的规律性; 2. 在时间上, 数据是随机的, 即数据在时间上的分布是随机的, 没有明显的规律性; 3. 在空间和时间上, 数据是随机的, 即数据在空间和时间上的分布是随机的, 没有明显的规律性。

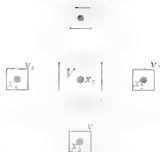
10.2.4 普通克立格插值

普通克立格插值是一种基于空间统计学的插值方法, 它假设数据在空间上是随机的, 且数据在空间上的分布是均匀的。普通克立格插值的核心思想是: 在已知数据点的基础上, 通过计算未知点与已知数据点的距离, 来估计未知点的值。普通克立格插值的具体步骤如下: 1. 计算已知数据点的平均值; 2. 计算未知点与已知数据点的距离; 3. 根据距离的大小, 对未知点的值进行估计。

普通克立格插值是一种基于空间统计学的插值方法, 它假设数据在空间上是随机的, 且数据在空间上的分布是均匀的。普通克立格插值的核心思想是: 在已知数据点的基础上, 通过计算未知点与已知数据点的距离, 来估计未知点的值。普通克立格插值的具体步骤如下: 1. 计算已知数据点的平均值; 2. 计算未知点与已知数据点的距离; 3. 根据距离的大小, 对未知点的值进行估计。

普通克立格插值是一种基于空间统计学的插值方法, 它假设数据在空间上是随机的, 且数据在空间上的分布是均匀的。普通克立格插值的核心思想是: 在已知数据点的基础上, 通过计算未知点与已知数据点的距离, 来估计未知点的值。普通克立格插值的具体步骤如下: 1. 计算已知数据点的平均值; 2. 计算未知点与已知数据点的距离; 3. 根据距离的大小, 对未知点的值进行估计。

普通克立格插值是一种基于空间统计学的插值方法, 它假设数据在空间上是随机的, 且数据在空间上的分布是均匀的。普通克立格插值的核心思想是: 在已知数据点的基础上, 通过计算未知点与已知数据点的距离, 来估计未知点的值。普通克立格插值的具体步骤如下: 1. 计算已知数据点的平均值; 2. 计算未知点与已知数据点的距离; 3. 根据距离的大小, 对未知点的值进行估计。

图 10-2 $n=4$ 时信息样点和待估块段承载图(或估计构形图)

将估计值 $\hat{Z}_V = \frac{1}{V} \sum_{i=1}^n \lambda_i Z(x_i)$ 代入 $E(Z_V - Z_V) = 0$ 中, 解得 λ_i 的估计值 $\hat{\lambda}_i$ 。在满足无偏性条件下, 求出 n 个权系数 λ_i 。

10.2.4.2 普通克立格法

普通克立格法与普通克里金法类似, 只是将 $Z(x_i)$ 替换为 $Z(x_i)$ 。在满足无偏性条件下, 求出 n 个权系数 λ_i 。将估计值 $\hat{Z}_V = \frac{1}{V} \sum_{i=1}^n \lambda_i Z(x_i)$ 代入 $E(Z_V - Z_V) = 0$ 中, 解得 λ_i 的估计值 $\hat{\lambda}_i$ 。

$$E(Z_V - Z_V) = 0 \quad (10.41)$$

$$\text{因为 } E(Z_V) = \frac{1}{V} \int_V E[Z(x)] dx = \mu$$

$$\hat{Z}_V = \frac{1}{V} \sum_{i=1}^n \lambda_i Z(x_i) = \frac{1}{V} \sum_{i=1}^n \lambda_i \mu = \mu \sum_{i=1}^n \lambda_i$$

故得无偏性条件:

$$\sum_{i=1}^n \lambda_i = 1 \quad (10.42)$$

在满足无偏性条件下, 估计方差 σ_E^2 为:

$$\begin{aligned} \sigma_E^2 &= E(Z_V^2) - \mu^2 = E\left(\left(\frac{1}{V} \sum_{i=1}^n \lambda_i Z(x_i)\right)^2\right) - \mu^2 \\ &= \frac{1}{V^2} \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j E[Z(x_i) Z(x_j)] - \mu^2 \end{aligned} \quad (10.43)$$

根据克里金原理, 令:

$$F = \sigma_E^2 = \frac{1}{V^2} \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j C(x_i, x_j) - \mu^2 \quad (10.44)$$

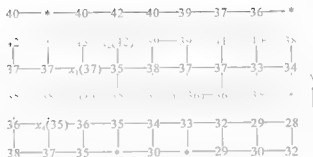
图 10-3 空间正方形网格数据(点间距 $h=1$ km)

图 10-4 缺失值情况下样本数对的组成和计算过程 为缺失值,

$$\begin{aligned}
 \gamma(1) &= \frac{1}{2 \times 36} [(38-35)^2 + (35-37)^2 + (40-43)^2 + (43-37)^2 + (36-35)^2 + (42-42)^2 + \\
 &\quad (42-35)^2 + (35-35)^2 + (35-35)^2 + (40-39)^2 + (39-38)^2 + (38-37)^2 + \\
 &\quad (37-34)^2 + (34-30)^2 + (39-39)^2 + (39-37)^2 + (37-36)^2 + (36-33)^2 + \\
 &\quad (37-41)^2 + (41-37)^2 + (37-36)^2 + (36-32)^2 + (32-29)^2 + (36-40)^2 + \\
 &\quad (40-33)^2 + (33-35)^2 + (35-29)^2 + (29-30)^2 + (38-34)^2 + (28-32)^2] \\
 &= 385/72 = 5.35 \\
 \gamma(2) &= \frac{1}{2 \times 27} [(40-37)^2 + (42-35)^2 + (37-36)^2 + (35-38)^2 + (37-35)^2 + (38-37)^2 + \\
 &\quad (40-37)^2 + (37-36)^2 + (42-35)^2 + (42-35)^2 + (35-35)^2 + (40-38)^2 + \\
 &\quad (39-37)^2 + (38-34)^2 + (37-30)^2 + (39-37)^2 + (39-36)^2 + (37-33)^2 + \\
 &\quad (37-37)^2 + (41-36)^2 + (37-32)^2 + (36-29)^2 + (36-33)^2 + (40-35)^2 + \\
 &\quad (33-29)^2 + (35-30)^2 + (34-28)^2] = 493/54 = 9.13
 \end{aligned}$$

$$\begin{aligned} & (42-35)^2 + (42-35)^2 + (40-37)^2 + (39-34)^2 + (38-30)^2 + (39-36)^2 + \\ & (39-33)^2 + (37-36)^2 + (41-32)^2 + (37-29)^2 + (36-35)^2 + (40-29)^2 + \\ & (33-30)^2 + (38-28)^2 + (34-32)^2 = 737/42 = 17.55 \end{aligned}$$

$$\frac{(39-30)^2 + (39-33)^2 + (37-32)^2 + (41-29)^2 + (36-29)^2 + (40-30)^2 + (38-32)^2}{7} = 668/7 = 25.69$$

229/10 -22.90

[illegible]

4. 月利率为 0.01, 计算 10 年, 10 年后, 月利率变为 0.02, 其他日不变, 结果 (表 10.2)。

表 10.2 南北、西北—东南方向上的变差函数计算结果

| 方向 | | 南北 | | | | 方向 | | | | 西北 | 东南 |
|-------------|------|------|-------|-------|-------|-------------|------|-------|-------|-------|-------|
| h | 1 | 2 | 3 | 4 | 5 | h | 1.41 | 2.82 | 4.24 | 5.65 | 7.07 |
| $N(h)$ | 36 | 27 | 21 | 13 | 5 | $N(h)$ | 32 | 21 | 13 | 8 | 2 |
| $\gamma(h)$ | 5.35 | 9.13 | 17.55 | 25.69 | 22.90 | $\gamma(h)$ | 7.06 | 12.95 | 30.85 | 58.13 | 50.00 |

从上述两个定理可以推出, 我们可以在球面上找到有限个一般形人力,

0 ($h=0$)

$$c_0 \neq c \quad (h > a)$$

当 $0 < h < a$ 时, 有:

由式(10.54)可知, 当 $0 < h < a$ 时, 球状变差函数模型为: $\gamma(h) = b_0 + b_1x_1 + b_2x_2$, 其中, $x_1 = h/a$, $x_2 = (h/a)^2$, 且, 线性模型:

$$y = b_0 + b_1x_1 + b_2x_2 \quad (10.57)$$

在式(10.57)中, 代入数据, 用最小二乘法求得, 可得:

$$y = 2.048 + 1.731x_1 - 0.00792x_2 \quad (10.58)$$

由式(10.58)可知, 球状变差函数模型中, 系数 $b_1 = 1.731$, $b_2 = -0.00792$, $R^2 = 0.999$, 可见模型的拟合效果是很好的。

由式(10.54)可知, 当 $h > a$ 时, 球状变差函数模型为: $\gamma(h) = a^2$, 其中, $a = 8.353$, 所以, 球状变差函数模型为:

$$\gamma(h) = 0 \quad (h=0)$$

$$\gamma(h) = 3.202 \quad (h > 8.353)$$

在式(10.59)中, 代入数据, 求得拟合模型, 拟合效果很好, 有关软件未完成。

例 10.2 克立格估计实例(徐建华, 2006)

某地区有 4 个采样点, 其坐标如图 10-1 所示, 已知这 4 个采样点的污染指数 $Z(x_i)$ 分别为 10, 15, 20, 25, 求该地区的污染指数 $Z(x_0)$ 。解: 由图 10-1 可知, 4 个采样点的坐标分别为 $(x_1, y_1) = (1, 1)$, $(x_2, y_2) = (2, 2)$, $(x_3, y_3) = (3, 3)$, $(x_4, y_4) = (4, 4)$, 且, 污染指数 $Z(x_i)$ 分别为 10, 15, 20, 25。由式(10.55)可知, 污染指数 $Z(x_0)$ 的估计值为:

$$Z^*(x_0) = \sum_{i=1}^4 \lambda_i Z(x_i) \quad (i = 1, 2, 3, 4)$$

根据式(10.55), 可知:

$$\begin{aligned} \lambda_1 &= \begin{vmatrix} c_{11} & c_{12} & c_{13} & c_{14} & 1 \\ c_{21} & c_{22} & c_{23} & c_{24} & 1 \\ c_{31} & c_{32} & c_{33} & c_{34} & 1 \\ c_{41} & c_{42} & c_{43} & c_{44} & 1 \end{vmatrix} / \begin{vmatrix} c_{11} & c_{12} & c_{13} & c_{14} & 1 \\ c_{21} & c_{22} & c_{23} & c_{24} & 1 \\ c_{31} & c_{32} & c_{33} & c_{34} & 1 \\ c_{41} & c_{42} & c_{43} & c_{44} & 1 \end{vmatrix} \\ \lambda_2 &= \begin{vmatrix} c_{11} & c_{12} & c_{13} & c_{14} & 1 \\ c_{21} & c_{22} & c_{23} & c_{24} & 1 \\ c_{31} & c_{32} & c_{33} & c_{34} & 1 \\ c_{41} & c_{42} & c_{43} & c_{44} & 1 \end{vmatrix} / \begin{vmatrix} c_{11} & c_{12} & c_{13} & c_{14} & 1 \\ c_{21} & c_{22} & c_{23} & c_{24} & 1 \\ c_{31} & c_{32} & c_{33} & c_{34} & 1 \\ c_{41} & c_{42} & c_{43} & c_{44} & 1 \end{vmatrix} \\ \lambda_3 &= \begin{vmatrix} c_{11} & c_{12} & c_{13} & c_{14} & 1 \\ c_{21} & c_{22} & c_{23} & c_{24} & 1 \\ c_{31} & c_{32} & c_{33} & c_{34} & 1 \\ c_{41} & c_{42} & c_{43} & c_{44} & 1 \end{vmatrix} / \begin{vmatrix} c_{11} & c_{12} & c_{13} & c_{14} & 1 \\ c_{21} & c_{22} & c_{23} & c_{24} & 1 \\ c_{31} & c_{32} & c_{33} & c_{34} & 1 \\ c_{41} & c_{42} & c_{43} & c_{44} & 1 \end{vmatrix} \\ \lambda_4 &= \begin{vmatrix} c_{11} & c_{12} & c_{13} & c_{14} & 1 \\ c_{21} & c_{22} & c_{23} & c_{24} & 1 \\ c_{31} & c_{32} & c_{33} & c_{34} & 1 \\ c_{41} & c_{42} & c_{43} & c_{44} & 1 \end{vmatrix} / \begin{vmatrix} c_{11} & c_{12} & c_{13} & c_{14} & 1 \\ c_{21} & c_{22} & c_{23} & c_{24} & 1 \\ c_{31} & c_{32} & c_{33} & c_{34} & 1 \\ c_{41} & c_{42} & c_{43} & c_{44} & 1 \end{vmatrix} \end{aligned} \quad (10.60)$$

由式(10.60)可知, 污染指数 $Z(x_0)$ 的估计值为:

$$3.202$$

$$(h=0)$$

$$c^*(h) = \begin{cases} 1.154 \left[1 - \left(\frac{3h}{2 \times 8.535} - \frac{h^2}{2 \times 8.535^2} \right) \right] & (0 < h \leq 8.535) \\ 0 & (h > 8.535) \end{cases}$$

$$\text{当 } i=j \text{ 时, } c_{11}=c_{22}=c_{33}=c_{44}=c(0)=c_1+c_2=2.048+1.154=3.202$$

$$c_{12}=c_{21}=c_{13}=c_{31}=c_{14}=c_{41}=c_{24}=c_{42}=c_{34}=c_{43}=c(h)=c_1+c_2=2.048+1.154=3.202$$

由此计算可得:

$$c_{11}=c_{22}=c_{33}=c_{44}=c(0)=c_1+c_2=2.048+1.154=3.202$$

$$c_{12}=c_{21}=c_{13}=c_{31}=c_{14}=c_{41}=c_{24}=c_{42}=c_{34}=c_{43}=c(h)=c_1+c_2=2.048+1.154=3.202$$

$$c_{23}=c_{32}=3.202 - \gamma(\sqrt{2^2+2^2})=0.601$$

$$c_{34}=c_{43}=3.202 - \gamma(\sqrt{4^2+1^2})=0.383$$

$$c_{01}=3.202 - \gamma(\sqrt{1^2})=0.952$$

将

将以上计算结果代入克立格方程组(10.54), 得:

$$\lambda_1 | \begin{bmatrix} 3.202 & 0.870 & 0.542 & 0.711 & 1.000 \end{bmatrix}^{-1} \begin{bmatrix} 0.952 \end{bmatrix} = \begin{bmatrix} 0.287 \end{bmatrix}$$

$$\lambda_2 | \begin{bmatrix} 0.870 & 3.202 & 0.601 & 0.466 & 1.000 \end{bmatrix} \begin{bmatrix} 0.711 \end{bmatrix} = \begin{bmatrix} 0.210 \end{bmatrix}$$

$$\lambda_3 = \begin{bmatrix} 0.542 & 0.601 & 3.202 & 0.383 & 1.000 \end{bmatrix} \begin{bmatrix} 0.571 \end{bmatrix} = \begin{bmatrix} 0.202 \end{bmatrix}$$

$$\lambda_4 = \begin{bmatrix} 0.711 & 0.466 & 0.383 & 3.202 & 1.000 \end{bmatrix} \begin{bmatrix} 0.870 \end{bmatrix} = \begin{bmatrix} 0.301 \end{bmatrix}$$

$$\mu = \begin{bmatrix} 1.000 & 1.000 & 1.000 & 1.000 & 0.000 \end{bmatrix} \begin{bmatrix} 1.000 \end{bmatrix} = \begin{bmatrix} -0.473 \end{bmatrix}$$

于是, 格点 x_0 处降水量的克立格估计值为:

$\mu = -0.473$, 所以 x_0 点的降水量的克立格估计值为:

$$Z_0^* = 0.287Z(x_1) + 0.210Z(x_2) + 0.202Z(x_3) - 0.301Z(x_4)$$

$$= 0.287 \times 37 + 0.210 \times 42 + 0.202 \times 36 - 0.301 \times 35 = 37.250(\text{mm})$$

克立格估计方差为:

$$\sigma_K^2 = c(x_0, x_0) - \sum_{i=1}^4 \lambda_i c(x_i, x_0) + \mu$$

$$= 3.202 - (0.287 \times 0.952 + 0.210 \times 0.711 + 0.202 \times 0.571 + 0.301 \times 0.870) + 0.473$$

$$= 2.875(\text{mm})$$

例 10.10 某地区有 5 个气象站, 测得 1980 年 1 月 1 日的气温数据如下:

$x_1 = 15.2, x_2 = 18.5, x_3 = 20.1, x_4 = 22.3, x_5 = 24.7$ (单位: $^{\circ}\text{C}$)

在 ArcGIS 中, 将采样点数据加载到 ArcMap 中, 如图 10-5 所示, 采样点分布图。

例 10.3 软件计算实例

(1) 数据采集

在 ArcGIS 中, 将采样点数据加载到 ArcMap 中, 如图 10-5 所示, 采样点分布图。在 ArcMap 中, 将采样点数据加载到 ArcMap 中, 如图 10-5 所示, 采样点分布图。在 ArcMap 中, 将采样点数据加载到 ArcMap 中, 如图 10-5 所示, 采样点分布图。

简称 SD) 三种水质参数



图 10-5 采样点分布图

(2) 异常值的识别与处理 影响系数法

在 ArcGIS 中, 将采样点数据加载到 ArcMap 中, 如图 10-5 所示, 采样点分布图。在 ArcGIS 中, 将采样点数据加载到 ArcMap 中, 如图 10-5 所示, 采样点分布图。在 ArcGIS 中, 将采样点数据加载到 ArcMap 中, 如图 10-5 所示, 采样点分布图。在 ArcGIS 中, 将采样点数据加载到 ArcMap 中, 如图 10-5 所示, 采样点分布图。在 ArcGIS 中, 将采样点数据加载到 ArcMap 中, 如图 10-5 所示, 采样点分布图。

表 10.3

影响系数法计算结果

| 采样点 Z_i | Z_{12} | Z_{13} | Z_{14} | Z_{15} | Z_{16} | Z_{17} | Z_{18} | Z_{19} | Z_{20} |
|--------------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
| SS 31 | 28 | 44 | 28 | 25 | 24 | 18 | 17 | 30 | 27 |
| M | 27.55 | 27.55 | 27.55 | 27.55 | 27.55 | 27.55 | 27.55 | 27.55 | 27.55 |
| M/m | 1.006 6 | 1.000 9 | 1.032 5 | 1.000 9 | 0.996 2 | 0.993 3 | 0.982 1 | 0.980 2 | 1.004 7 |
| 采样点 Z_{2i} | Z_{22} | Z_{23} | Z_{24} | Z_{25} | Z_{26} | Z_{27} | Z_{28} | Z_{29} | Z_{30} |
| SS 38 | 30 | 22 | 26 | 12 | 24 | 26 | 20 | 38 | 43 |
| M | 27.55 | 27.55 | 27.55 | 27.55 | 27.55 | 27.55 | 27.55 | 27.55 | 27.55 |
| m | 27.000 0 | 27.421 1 | 27.842 1 | 27.631 6 | 28.368 1 | 27.736 8 | 27.631 6 | 27.947 4 | 27.000 0 |
| M/m | 1.020 4 | 1.004 7 | 0.989 5 | 0.997 1 | 0.971 2 | 0.993 3 | 0.997 1 | 0.985 8 | 1.020 4 |

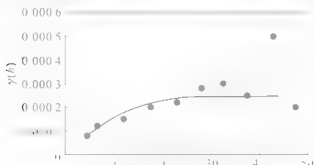
1. 计算 M 和 m 值。由表 10.3 可知, 当 $k=0.1$ 时, $M=27.55$, $m=27.000 0$; 当 $k=0.05$ 时, $M=27.55$, $m=27.421 1$ 。由表 10.3 可知, 当 $k=0.1$ 时, $M/m=1.006 6$; 当 $k=0.05$ 时, $M/m=1.000 9$ 。由表 10.3 可知, 当 $k=0.1$ 时, $M/m=1.006 6$; 当 $k=0.05$ 时, $M/m=1.000 9$ 。

值。本例 $k=0.1$ 时, $GL=75.14$; $k=0.05$ 时, $GL=52.48$ 。

由表 10.3 可知, 当 $k=0.1$ 时, $M/m=1.006 6$; 当 $k=0.05$ 时, $M/m=1.000 9$ 。

(3) 实验变差函数

由表 10.3 可知, 当 $k=0.1$ 时, $M/m=1.006 6$; 当 $k=0.05$ 时, $M/m=1.000 9$ 。



叶绿素 a 的变差图

表 10.4 四种变差函数理论模型拟合参数表

| 理论模型 | c_0 | c | a | Q | S | R^2 |
|----------|-----------|----------|----------|-------------|----------|---------|
| 线性有基台值模型 | 100.164 0 | 50.028 0 | 44.893 0 | 1 301.251 9 | 12.753 7 | 0.655 2 |
| 球状模型 | 100.107 9 | 35.114 0 | 33.829 5 | 375.624 1 | 6.852 2 | 0.747 0 |
| 指数模型 | 90.483 0 | 57.528 0 | 51.325 0 | 2 359.848 9 | 17.175 0 | 0.542 3 |
| 高斯模型 | 104.815 0 | 50.372 0 | 52.013 0 | 1 453.124 9 | 13.477 4 | 0.516 8 |

球状模型理论变差函数拟合结果。

表 10.5 水质参数理论变差函数拟合结果(球状模型拟合)

| | c_0 | c | a | Q | S | R^2 |
|-------|-----------|-----------|-------------|-----|-----|----------|
| chl-a | 0.000 02 | 0.000 22 | 0.000 24 | | | 29.152 1 |
| SS | 100.107 9 | 35.114 0 | 135.221 9 | | | 33.829 5 |
| SD | 200.013 5 | 850.201 9 | 1 050.215 4 | | | 35.171 2 |

(5) 空间最优估计

表 10.6。

表 10.6 克立格内插与传统估计方法结果比较

| 站号 | 采样日期 | 估计值 | 估计误差 | 估计值 | 估计误差 | 估计值 | 估计误差 |
|--------|------|-------|-------|-------|-------|-------|-------|
| 8 | 26 | 31.50 | -5.50 | 31.54 | -5.54 | 27.96 | -1.96 |
| 9 | 26 | 29.25 | 3.25 | 29.36 | -3.36 | 26.66 | -0.66 |
| 14 | 25 | 26.50 | -1.50 | 26.46 | -1.46 | 26.15 | -1.15 |
| 15 | 24 | 23.75 | 0.25 | 23.56 | 0.44 | 24.09 | -0.09 |
| 19 | 27 | 29.75 | -2.75 | 30.11 | -3.11 | 29.66 | 2.66 |
| 26 | 26 | 23.00 | 3.00 | 22.92 | 3.08 | 23.00 | 3.00 |
| 32 | 30 | 30.00 | 0.00 | 30.34 | -0.34 | 29.63 | 0.37 |
| 33 | 26 | 25.75 | 0.25 | 25.90 | 0.10 | 26.30 | -0.30 |
| 34 | 25 | 25.50 | 0.50 | 25.29 | 0.29 | 26.53 | -1.53 |
| 45 | 30 | 30.50 | -0.50 | 31.18 | 1.18 | 29.67 | 0.33 |
| 估计误差均值 | | -1.05 | | -1.17 | | 0.47 | |

[illegible]

(6) 太湖水质评价

$\{x_1, x_2, \dots, x_n\}$ 是 \mathbb{R}^n 中 n 个线性无关的向量, 且 $\{x_1, x_2, \dots, x_n, x_{n+1}\}$ 是 \mathbb{R}^{n+1} 中 $n+1$ 个线性无关的向量, 则 x_{n+1} 与 x_1, x_2, \dots, x_n 线性无关.



图 10-7 太湖中总悬浮物的评价结果

10.3 环境空间主成分分析

人分作三类,人数各不相同。在这些人中,有一个人说真话,有一个人说假话,有一个人说真话和假话的概率各为 $\frac{1}{2}$ 。现在,你问他们 N 个问题,问第 i 个问题,这个人回答是“是”或“否”。问完 N 个问题后,你根据 N 个人的回答,作出一个判断。问:你最多能问几个问题,使得你作出判断的概率大于 $\frac{1}{2}$?

主成分分析需经过以下主要步骤:

、各参数指标、利用、我、化等

处理方法对数据作相应变换:

参评因子的一般标准化量化公式为:

$$Q_i = \frac{\lambda_i - \lambda_{\min}}{X_{\max} - X_{\min}} \times 10 \quad (i=1, 2, \dots, N) \quad (10.61)$$

上式, Q_i 为参评因子 i 的无量纲化得分, λ_i 为参评因子第 i 级评价指标得分, λ_{\min} 为参评因子最低评价指标得分, X_{\max} 为参评因子最大评价指标得分。

(2) 建立 N 个变量的相关系数矩阵 R ;

其中 $R = (r_{ij})_{N \times N}$, R 为 N 阶实对称正定矩阵, 特征值 λ_i 满足:

(4) 将特征向量作线性组合, 输出 m 个主成分。

10.3.1 空间主成分分析步骤

环境空间主成分分析, 是指对数据为基础, 对数据进行数学的变换和旋转, 将数据由原来的 N 个主成分, 转换为 m 个主成分, 实现数据降维。主成分分析是多元统计分析中最重要的方法之一。主成分分析在环境科学中应用广泛, 如环境综合评价、环境风险评估、环境质量管理等。主成分分析的基本步骤如下:

- (1) 数据标准化: 将原始数据 $X = (x_{ij})_{n \times p}$ 进行标准化处理, 得到标准化数据 $Z = (z_{ij})_{n \times p}$ 。
- (2) 计算协方差矩阵: 计算标准化数据 Z 的协方差矩阵 $S = (s_{ij})_{p \times p}$ 。
- (3) 求解特征值和特征向量: 求解协方差矩阵 S 的特征值 $\lambda_1, \lambda_2, \dots, \lambda_p$ 和特征向量 v_1, v_2, \dots, v_p 。
- (4) 选择主成分: 根据特征值的大小, 选择前 m 个主成分。
- (5) 计算主成分得分: 将原始数据 X 投影到主成分空间, 得到主成分得分 $F = (f_{ij})_{n \times m}$ 。

在 APO-ENFO 空间主成分分析中, 将数据标准化为标准分数:

将数据标准化为标准分数, 其计算公式为:

将数据标准化为标准分数, 其计算公式为:

个综合图:

将数据标准化为标准分数, 其计算公式为:

$$a_i = \lambda_i / \sum_{j=1}^m \lambda_j \quad (i=1, 2, \dots, m) \quad (10.62)$$

上式得到各主成分的特征值, 根据特征值的大小, 确定主成分数:

将数据标准化为标准分数, 其计算公式为:

$$E = a_1 Y_1 + a_2 Y_2 + \dots + a_m Y_m \quad (j=1, 2, \dots, M) \quad (10.63)$$

其中, E 为环境综合评价指数; Y_j 为第 j 个主成分; a_j 为第 j 个主成分对应的贡献率。

从表 10.8 可以看出,在最终生成的生态环境综合评价数中,由于生和、水、土、气四因子权重所占权重分别为 0.28、0.28、0.28、0.16,基本上保留了原来变量反映的主要信息,可信度较高。

● 评价结果及其空间分布

在地理信息系统 ArcGIS 支持下,根据表 10.8 和式 (10.1),可以生成到了生态环境综合评价等级图(图 10.1),图中所显示如图表 10.9。根据生成生态环境综合评价图,将大宁河流域划分为 10 个不同的生态环境等级区。

表 10.9 大宁河流域生态环境综合评价结果

| 等级 | 面积/km ² | 面积比例 | 等级 | 面积/km ² | 面积比例 |
|----|--------------------|-------|----|--------------------|-------|
| 1 | 290.881 | 0.066 | 6 | 160.975 | 0.036 |
| 2 | 254.532 | 0.058 | 7 | 402.793 | 0.091 |
| 3 | 397.069 | 0.090 | 8 | 583.945 | 0.132 |
| 4 | 459.954 | 0.104 | 9 | 648.656 | 0.147 |
| 5 | 146.955 | 0.033 | 10 | 768.248 | 0.174 |
| 6 | 212.926 | 0.048 | 合计 | 718.238 | 0.163 |
| 7 | 627.912 | 0.142 | | 655.804 | 0.149 |
| 8 | 61.588 | 0.014 | | 374.530 | 0.085 |
| 9 | 1 743.096 | 0.395 | | 94.957 | 0.022 |
| 10 | 220.929 | 0.050 | | 7.697 | 0.002 |
| 合计 | 4 415.843 | 1.000 | | 4 415.843 | 1.000 |



1990 年

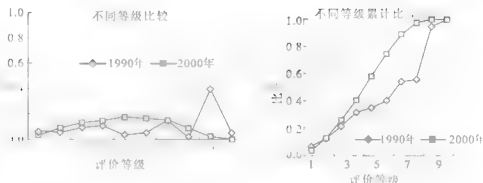


图 10-10 1990 年和 2000 年生态环境综合评价等级对比图

【思考题 10】

1. 什么是区域化变量? 什么是协方差函数和变差函数?
2. 什么是克立格方法? 举例说明克立格方法在环境科学中的应用。

【解答】(1) 区域化变量: 是指随空间位置而变化的变量, 如土壤肥力、降雨量、污染浓度等。

$$\begin{array}{ccccccc}
 15 & \text{---} & x_0(13) & \text{---} & 10 & \text{---} & 12 & \text{---} & 17 \\
 | & & | & & | & & | & & | \\
 18 & \text{---} & 15 & \text{---} & 11 & \text{---} & 14 & \text{---} & 19 \\
 \\
 16 & \text{---} & x_1(20) & \text{---} & x_0 & \text{---} & 15 & \text{---} & 23 \\
 | & & | & & | & & | & & | \\
 15 & \text{---} & 17 & \text{---} & 18 & \text{---} & x_1(18) & \text{---} & 21 \\
 \\
 10 & \text{---} & 16 & \text{---} & x_2(21) & \text{---} & 16 & \text{---} & 18
 \end{array}$$

4. 估计第 3 题中球状模型的参数。
5. 试用普通克立格估计第 3 题中 x_i 点的值。
6. 什么是主成分分析? 什么是空间主成分分析?
7. 空间主成分分析的步骤有哪些?
8. 空间主成分分析的目的是什么?
9. 在应用 ARC/INFO 进行空间主成分分析时, 都涉及到什么操作?
10. 举例说明空间主成分分析的作用。

部分思考题答案

思考题 1

4. 答案: $\mu = 1, \sigma = 1$. 因为 X 服从标准正态分布, 所以 X 的各阶中心矩等。

6. 答案: $\mu = 2, \sigma = 3$

$$\begin{aligned} P(3 < X < 9) &= \Phi\left(\frac{9-\mu}{\sigma}\right) - \Phi\left(\frac{3-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{9-2}{3}\right) - \Phi\left(\frac{3-2}{3}\right) = \Phi(2.333\ 3) - \Phi(0.333\ 3) \\ &= 0.990 = 0.629 - 0.361 \end{aligned}$$

7. 答案: $E\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i E(X_i) = \sum_{i=1}^n a_i \mu = \mu$

8. 答案: $n = 9, \mu_0 = 15, \mu_0 = 500, \alpha = 0.01$

$$\bar{x} = \sum_{i=1}^n x_i / n = 510.111\ 1$$

$$Z = \frac{\bar{x} - \mu_0}{\sigma_0 / \sqrt{n}}$$

$$z_{0.005} = 2.575\ 8$$

$$|Z| < z_{0.005} = 2.575\ 8$$

不能拒绝原假设。这台包装机工作正常。

9. 答案: $n = 7, \mu_0 = 300, \alpha = 0.05$

$$\bar{x} = \sum_{i=1}^n x_i / n = 296.285\ 7$$

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} = -0.359\ 6$$

$$t_{0.025}(6) = 2.447$$

$$|t| < t_{0.025}(6) = 2.447$$

不能拒绝原假设。灌溉对玉米穗重无显著影响。

10. 答案:

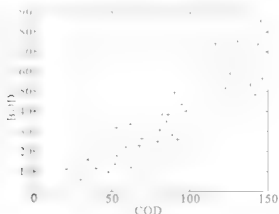
双因素方差分析表

| 方差来源 | SS | 自由度 | 均方差 | F 值 | |
|------|---------|-----|---------|----------|------------------|
| 因素 A | 0.157 4 | 2 | 0.078 7 | 23.826 7 | $F(2, 6) = 5.14$ |
| 因素 B | 0.879 6 | 3 | 0.293 2 | 88.756 5 | $F(3, 6) = 4.76$ |
| 误差 E | 0.019 8 | 6 | 0.003 3 | | |
| 总和 T | 1.056 8 | 11 | | | |

思考题 2

4. 答案:

(1) 根据表中所给的数据, 可以作出如下的散点图:



(4 题图 1)

可知 COD 与 BOD 之间存在线性关系。

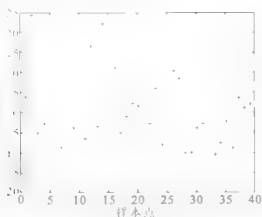
(3) 得到的线性回归方程为: $\hat{y} = a + bx = -5.364\ 2 + 0.492\ 5x$

(4) 因为 $L_{xx} = 5.450\ 3 \times 10^4$, $L_{xy} = 2.684\ 3 \times 10^4$, $L_{yy} = 1.685\ 5 \times 10^4$,

$$r = \frac{L_{xy}}{L_{xx} L_{yy}} = \frac{2.684\ 3 \times 10^4}{\sqrt{5.450\ 3 \times 10^4 \times 1.685\ 5 \times 10^4}} = 1.492\ 7$$

由此可知 COD 与 BOD 的决定系数 $r^2 = 2.228\ 3$ 。

(5) 根据残差的定义, 可以得到以下的残差图:



(4 题图 2)

$$\hat{y} = \bar{a} + bx = -5.364\ 2 + 0.492\ 5 \times 99 = 43.393\ 0$$

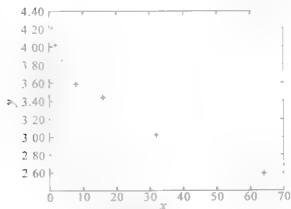
$$s^2 = \frac{Q}{n-2} = 0.280\ 1 \times 10^4$$

近似有

$$P(43.393\ 0 - 105.85 < y_0 < 43.393\ 0 + 105.85) = 0.95$$

5. 答案:

(1) 散点图:

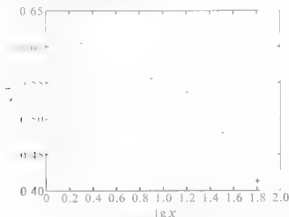


(5 题图 1)

(2) 作变换, 对 $y = ax^b$ 两边取对数, 得到 $\lg y = \lg a + b \lg x$

令 $y' = \lg y$, $x' = \lg x$, 则得 $y' = \lg a + bx'$

对数据进行变换, 变换后数据的散点图如下所示:



(5 题图 2)

根据变换后的数据, 采用一元线性回归分析可以得到:

$$\hat{y}' = \hat{A} + \hat{B}x' = 0.6429 - 0.1107x'$$

将 $y' = \lg y$, $x' = \lg x$ 代入到上式中可以得到:

$$y = 4.2160x^{-0.1107}$$

(3) $S_B = 1.77 - 0.0333$, $S_{\text{总}} = 0.0311$, $S_{\text{回}} = S_B - S_{\text{回}} = 0.0333 - 0.0311 = 0.0022$

$$F = \frac{S_{\text{回}}}{S_{\text{总}} / (n-2)} = 71.0158$$

认为线性关系式 $y' = A + Bx' = 0.6429 - 0.1107x'$ 显著。

思考题 3

2. 答案: 回归方程为: $y = 6.0944 - 0.0371x_1 + 0.0204x_2 - 1.1762x_3$ 。

3. 答案: 根据题意, 设 x_1 为产量, x_2 为成本, x_3 为利润, y 为周转量。则各变量为:

回归方程为:

$$y = 1.5112 \times 10^4 - 0.239x_1 - 0.2533 \times 10^4 x_2 - 0.0081 \times 10^4 x_3 + 0.0005 \times 10^4 x_4$$

4. 答案:

(1) 相关系数矩阵为:

$$C = \begin{pmatrix} 1.0000 & 0.3355 \\ 0.3355 & 1.0000 \end{pmatrix}$$

(2) 设欧氏距离矩阵为 A 。

| | | | | | | | |
|---------|---------|---------|---------|---------|---------|---------|---------|
| 0.000 0 | 1.871 2 | 1.361 9 | 2.103 0 | 1.242 1 | 1.428 3 | 1.605 1 | 1.196 8 |
| 1.871 2 | 0.000 0 | 0.860 2 | 1.522 8 | 1.029 1 | 1.232 0 | 0.948 8 | 0.878 7 |
| 1.361 9 | 0.860 2 | 0.000 0 | 1.476 8 | 1.088 0 | 0.658 2 | 0.455 6 | 0.654 8 |
| 2.103 0 | 1.522 8 | 1.476 8 | 0.000 0 | 1.569 7 | 1.257 9 | 1.313 9 | 1.307 1 |
| 1.242 1 | 1.029 1 | 1.088 0 | 1.569 7 | 0.000 0 | 1.125 8 | 1.286 1 | 0.566 3 |
| 1.428 3 | 1.232 0 | 0.658 2 | 1.257 9 | 1.125 8 | 0.000 0 | 0.579 1 | 0.641 8 |
| 1.605 1 | 0.948 8 | 0.455 6 | 1.313 9 | 1.286 1 | 0.579 1 | 0.000 0 | 0.747 0 |
| 1.196 8 | 0.878 7 | 0.654 8 | 1.307 1 | 0.566 3 | 0.641 8 | 0.747 0 | 0.000 0 |

(3) 最短距离聚类图, 如下图所示。



(2 题图)

注: 监测年份轴中的 1~8 分别对应 1993~2000 年。

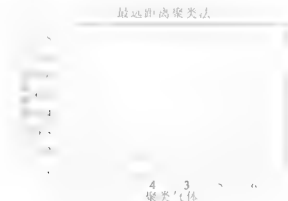
例 2 某地区 1993~2000 年 8 年来的 7 项环境指标数据如下表所示, 各指标均按 1993 年 = 100 进行无量纲化处理, 试按最短距离法进行聚类分析, 并画出最短距离聚类图。

为 A 。

| | | | | | | |
|-------|---------|---------|---------|---------|---------|---------|
| | 0.126 7 | 0.123 9 | 0.063 8 | 0.028 8 | 0.017 8 | 0.013 9 |
| | 0.110 9 | 0.081 1 | 0.205 8 | 0.083 5 | 0.048 3 | 0.004 6 |
| | 0.086 0 | 0.191 7 | 0.162 6 | 0.129 0 | 0.127 4 | 0.027 2 |
| X_1 | 0.076 9 | 0.140 1 | 0.119 3 | 0.121 4 | 0.439 5 | 0.018 4 |
| | 0.190 0 | 0.097 3 | 0.059 7 | 0.242 8 | 0.026 4 | 0.026 0 |
| | 0.144 8 | 0.106 2 | 0.205 8 | 0.159 3 | 0.061 5 | 0.873 8 |
| | 0.108 6 | 0.131 3 | 0.127 6 | 0.197 3 | 0.083 5 | 0.022 8 |
| | 0.156 1 | 0.128 3 | 0.055 6 | 0.037 9 | 0.195 6 | 0.013 3 |

| | | | | | |
|---------|---------|---------|---------|---------|---------|
| 1.000 0 | 0.897 5 | 0.821 1 | 0.873 7 | 0.563 6 | 0.441 8 |
| 0.897 5 | 1.000 0 | 0.872 0 | 0.836 6 | 0.736 5 | 0.344 6 |
| 0.821 1 | 0.872 0 | 1.000 0 | 0.816 6 | 0.601 0 | 0.566 2 |
| 0.873 7 | 0.836 6 | 0.816 6 | 1.000 0 | 0.552 6 | 0.443 3 |
| 0.563 6 | 0.736 5 | 0.601 0 | 0.552 6 | 1.000 0 | 0.158 5 |
| 0.441 8 | 0.344 6 | 0.566 2 | 0.443 3 | 0.158 5 | 1.000 0 |

最远距离聚类图，如下图所示。



(3 题图 1)

烷、环己烷为独立的气体，各为一类。

| | A | | | | | | B | | | | | |
|----------------|----------|----------|----------|---------|----------|----------|---|--|--|--|--|--|
| X ₂ | 0.048 2 | 0.035 5 | -1.047 3 | 1.379 8 | -0.820 0 | -0.392 3 | | | | | | |
| | 0.402 0 | -1.407 2 | 1.381 7 | 0.596 0 | -0.586 4 | -0.425 2 | | | | | | |
| | 1.109 6 | 2.140 4 | 0.642 5 | 0.057 2 | 0.018 7 | -0.345 4 | | | | | | |
| | -1.366 9 | 0.484 8 | -0.096 8 | 0.051 7 | 2.405 5 | -0.376 7 | | | | | | |
| | 1.849 3 | -0.886 9 | -1.117 7 | 1.690 1 | -0.754 5 | -0.349 8 | | | | | | |
| | 0.562 8 | -0.603 1 | 1.381 7 | 0.492 6 | -0.485 5 | 2.644 9 | | | | | | |
| | 0.466 3 | 0.201 0 | 0.044 0 | 1.036 9 | -0.317 5 | -0.361 0 | | | | | | |
| | 0.884 5 | 0.106 4 | 1.188 1 | 1.249 2 | 0.539 8 | 0.394 6 | | | | | | |

| | | | | | | | | |
|---|---------|---------|---------|---------|---------|---------|---------|---------|
| B | 0.000 0 | 2.941 9 | 3.420 7 | 3.917 2 | 3.661 2 | 4.396 6 | 2.757 9 | 1.614 1 |
| | 2.941 9 | 0.000 0 | 3.798 9 | 3.993 3 | 4.104 5 | 3.492 7 | 2.668 9 | 3.499 4 |
| | 3.420 7 | 3.798 9 | 0.000 0 | 3.010 5 | 4.927 7 | 4.500 7 | 2.367 8 | 3.666 7 |
| | 3.917 2 | 3.993 3 | 3.010 5 | 0.000 0 | 5.127 2 | 4.987 8 | 3.084 0 | 3.364 2 |
| | 3.661 2 | 4.104 5 | 4.927 7 | 5.127 2 | 0.000 0 | 4.296 2 | 2.917 7 | 3.498 5 |
| | 4.396 6 | 3.492 7 | 4.500 7 | 4.987 8 | 4.296 2 | 0.000 0 | 3.585 5 | 4.531 5 |
| | 2.757 9 | 2.668 9 | 2.367 8 | 3.084 0 | 2.917 7 | 3.585 5 | 0.000 0 | 3.051 9 |
| | 1.614 1 | 3.499 4 | 3.666 7 | 3.364 2 | 3.498 5 | 4.531 5 | 3.051 9 | 0.000 0 |

最短距离聚类图, 如下图所示。



(3题图2)

点分为两类, 则1和8号样点为一类; 2、3和7号为 一类。

4. 答案: 设标准差标准化后的矩阵为 X 。

| | | | | | | | |
|---|----------|---------|---------|---------|----------|---------|----------|
| X | 1.116 1 | 1.709 8 | 0.425 0 | 0.098 6 | 2.173 3 | 1.175 2 | 1.096 8 |
| | 2.616 9 | 0.854 0 | 0.655 9 | 0.713 0 | 0.530 8 | 0.745 6 | 0.620 6 |
| | -0.396 3 | 0.321 7 | 0.253 9 | 0.290 6 | 0.970 2 | 0.489 2 | 1.414 7 |
| | 0.606 5 | 0.805 0 | 0.597 5 | 0.668 0 | 0.000 1 | 0.608 4 | -0.776 8 |
| | 0.889 3 | 1.778 7 | 0.059 6 | 1.091 2 | 1.194 4 | 0.295 4 | 0.860 4 |
| | 0.423 7 | 0.405 5 | 0.782 6 | 0.552 0 | 1.023 6 | 0.955 7 | 1.133 6 |
| | 0.396 3 | 0.135 1 | 1.027 4 | 0.005 0 | 0.202 1 | 0.509 8 | 0.082 0 |
| | -0.045 3 | 0.321 6 | 1.245 7 | 1.130 3 | -0.516 7 | 2.101 4 | 0.851 5 |
| | 0.212 9 | 1.271 3 | 1.671 8 | 2.171 0 | 0.326 3 | 0.206 2 | 0.985 3 |
| | 0.515 0 | 0.019 2 | 1.396 4 | 0.671 3 | 0.406 5 | 0.857 1 | 1.038 8 |

欧氏距离矩阵为 A 。

| | | | | | | | | | |
|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| 0.000 0 | 5.009 3 | 2.142 6 | 4.410 4 | 5.603 9 | 5.098 2 | 3.419 1 | 3.394 6 | 3.209 2 | 4.739 4 |
| 5.009 3 | 0.000 0 | 4.497 4 | 3.505 6 | 3.787 8 | 3.477 3 | 3.893 8 | 4.747 0 | 4.870 5 | 3.862 8 |
| 2.142 6 | 4.497 4 | 0.000 0 | 3.050 9 | 4.141 7 | 3.752 9 | 2.016 4 | 2.863 7 | 3.035 5 | 3.478 3 |
| 4.410 4 | 3.505 6 | 3.050 9 | 0.000 0 | 1.780 8 | 1.239 6 | 1.730 4 | 4.300 7 | 4.702 8 | 1.218 0 |
| 5.603 9 | 3.787 8 | 4.141 7 | 1.780 8 | 0.000 0 | 1.908 5 | 3.223 1 | 4.541 8 | 5.458 6 | 2.549 6 |
| 5.098 2 | 3.477 3 | 3.752 9 | 1.239 6 | 1.908 5 | 0.000 0 | 2.060 4 | 4.601 9 | 4.935 3 | 0.973 6 |
| 3.419 1 | 3.893 8 | 2.016 4 | 1.730 4 | 3.223 1 | 2.060 4 | 0.000 0 | 3.811 3 | 3.804 2 | 1.596 2 |
| 3.394 6 | 4.747 0 | 2.863 7 | 4.300 7 | 4.541 8 | 4.601 9 | 3.811 3 | 0.000 0 | 2.560 4 | 4.787 1 |
| 3.209 2 | 4.870 5 | 3.035 5 | 4.702 8 | 5.458 6 | 4.935 3 | 3.804 2 | 2.560 4 | 0.000 0 | 5.049 2 |
| 4.739 4 | 3.862 8 | 3.478 3 | 1.218 0 | 2.549 6 | 0.973 6 | 1.596 2 | 4.787 1 | 5.049 2 | 0.000 0 |

最远距离聚类法



(4 题图)

3 和 7 号样点为一类。

5. 答案: 极差的标准化后的矩阵为 X , 指标间的相关系数为 R 。

1. 2. 3. 4. 5. 6. 7. 8. 9. 10. 11. 12. 13. 14. 15. 16. 17. 18. 19. 20. 21. 22. 23. 24. 25. 26. 27. 28. 29. 30. 31. 32. 33. 34. 35. 36. 37. 38. 39. 40. 41. 42. 43. 44. 45. 46. 47. 48. 49. 50. 51. 52. 53. 54. 55. 56. 57. 58. 59. 60. 61. 62. 63. 64. 65. 66. 67. 68. 69. 70. 71. 72. 73. 74. 75. 76. 77. 78. 79. 80. 81. 82. 83. 84. 85. 86. 87. 88. 89. 90. 91. 92. 93. 94. 95. 96. 97. 98. 99. 100. 101. 102. 103. 104. 105. 106. 107. 108. 109. 110. 111. 112. 113. 114. 115. 116. 117. 118. 119. 120. 121. 122. 123. 124. 125. 126. 127. 128. 129. 130. 131. 132. 133. 134. 135. 136. 137. 138. 139. 140. 141. 142. 143. 144. 145. 146. 147. 148. 149. 150. 151. 152. 153. 154. 155. 156. 157. 158. 159. 160. 161. 162. 163. 164. 165. 166. 167. 168. 169. 170. 171. 172. 173. 174. 175. 176. 177. 178. 179. 180. 181. 182. 183. 184. 185. 186. 187. 188. 189. 190. 191. 192. 193. 194. 195. 196. 197. 198. 199. 200. 201. 202. 203. 204. 205. 206. 207. 208. 209. 210. 211. 212. 213. 214. 215. 216. 217. 218. 219. 220. 221. 222. 223. 224. 225. 226. 227. 228. 229. 230. 231. 232. 233. 234. 235. 236. 237. 238. 239. 240. 241. 242. 243. 244. 245. 246. 247. 248. 249. 250. 251. 252. 253. 254. 255. 256. 257. 258. 259. 260. 261. 262. 263. 264. 265. 266. 267. 268. 269. 270. 271. 272. 273. 274. 275. 276. 277. 278. 279. 280. 281. 282. 283. 284. 285. 286. 287. 288. 289. 290. 291. 292. 293. 294. 295. 296. 297. 298. 299. 300. 301. 302. 303. 304. 305. 306. 307. 308. 309. 310. 311. 312. 313. 314. 315. 316. 317. 318. 319. 320. 321. 322. 323. 324. 325. 326. 327. 328. 329. 330. 331. 332. 333. 334. 335. 336. 337. 338. 339. 340. 341. 342. 343. 344. 345. 346. 347. 348. 349. 350. 351. 352. 353. 354. 355. 356. 357. 358. 359. 360. 361. 362. 363. 364. 365. 366. 367. 368. 369. 370. 371. 372. 373. 374. 375. 376. 377. 378. 379. 380. 381. 382. 383. 384. 385. 386. 387. 388. 389. 390. 391. 392. 393. 394. 395. 396. 397. 398. 399. 400. 401. 402. 403. 404. 405. 406. 407. 408. 409. 410. 411. 412. 413. 414. 415. 416. 417. 418. 419. 420. 421. 422. 423. 424. 425. 426. 427. 428. 429. 430. 431. 432. 433. 434. 435. 436. 437. 438. 439. 440. 441. 442. 443. 444. 445. 446. 447. 448. 449. 450. 451. 452. 453. 454. 455. 456. 457. 458. 459. 460. 461. 462. 463. 464. 465. 466. 467. 468. 469. 470. 471. 472. 473. 474. 475. 476. 477. 478. 479. 480. 481. 482. 483. 484. 485. 486. 487. 488. 489. 490. 491. 492. 493. 494. 495. 496. 497. 498. 499. 500. 501. 502. 503. 504. 505. 506. 507. 508. 509. 510. 511. 512. 513. 514. 515. 516. 517. 518. 519. 520. 521. 522. 523. 524. 525. 526. 527. 528. 529. 530. 531. 532. 533. 534. 535. 536. 537. 538. 539. 540. 541. 542. 543. 544. 545. 546. 547. 548. 549. 550. 551. 552. 553. 554. 555. 556. 557. 558. 559. 560. 561. 562. 563. 564. 565. 566. 567. 568. 569. 570. 571. 572. 573. 574. 575. 576. 577. 578. 579. 580. 581. 582. 583. 584. 585. 586. 587. 588. 589. 590. 591. 592. 593. 594. 595. 596. 597. 598. 599. 600. 601. 602. 603. 604. 605. 606. 607. 608. 609. 610. 611. 612. 613. 614. 615. 616. 617. 618. 619. 620. 621. 622. 623. 624. 625. 626. 627. 628. 629. 630. 631. 632. 633. 634. 635. 636. 637. 638. 639. 640. 641. 642. 643. 644. 645. 646. 647. 648. 649. 650. 651. 652. 653. 654. 655. 656. 657. 658. 659. 660. 661. 662. 663. 664. 665. 666. 667. 668. 669. 670. 671. 672. 673. 674. 675. 676. 677. 678. 679. 680. 681. 682. 683. 684. 685. 686. 687. 688. 689. 690. 691. 692. 693. 694. 695. 696. 697. 698. 699. 700. 701. 702. 703. 704. 705. 706. 707. 708. 709. 710. 711. 712. 713. 714. 715. 716. 717. 718. 719. 720. 721. 722. 723. 724. 725. 726. 727. 728. 729. 730. 731. 732. 733. 734. 735. 736. 737. 738. 739. 740. 741. 742. 743. 744. 745. 746. 747. 748. 749. 750. 751. 752. 753. 754. 755. 756. 757. 758. 759. 760. 761. 762. 763. 764. 765. 766. 767. 768. 769. 770. 771. 772. 773. 774. 775. 776. 777. 778. 779. 780. 781. 782. 783. 784. 785. 786. 787. 788. 789. 790. 791. 792. 793. 794. 795. 796. 797. 798. 799. 800. 801. 802. 803. 804. 805. 806. 807. 808. 809. 810. 811. 812. 813. 814. 815. 816. 817. 818. 819. 820. 821. 822. 823. 824. 825. 826. 827. 828. 829. 830. 831. 832. 833. 834. 835. 836. 837. 838. 839. 840. 841. 842. 843. 844. 845. 846. 847. 848. 849. 850. 851. 852. 853. 854. 855. 856. 857. 858. 859. 860. 861. 862. 863. 864. 865. 866. 867. 868. 869. 870. 871. 872. 873. 874. 875. 876. 877. 878. 879. 880. 881. 882. 883. 884. 885. 886. 887. 888. 889. 890. 891. 892. 893. 894. 895. 896. 897. 898. 899. 900. 901. 902. 903. 904. 905. 906. 907. 908. 909. 910. 911. 912. 913. 914. 915. 916. 917. 918. 919. 920. 921. 922. 923. 924. 925. 926. 927. 928. 929. 930. 931. 932. 933. 934. 935. 936. 937. 938. 939. 940. 941. 942. 943. 944. 945. 946. 947. 948. 949. 950. 951. 952. 953. 954. 955. 956. 957. 958. 959. 960. 961. 962. 963. 964. 965. 966. 967. 968. 969. 970. 971. 972. 973. 974. 975. 976. 977. 978. 979. 980. 981. 982. 983. 984. 985. 986. 987. 988. 989. 990. 991. 992. 993. 994. 995. 996. 997. 998. 999. 1000. 1001. 1002. 1003. 1004. 1005. 1006. 1007. 1008. 1009. 1010. 1011. 1012. 1013. 1014. 1015. 1016. 1017. 1018. 1019. 1020. 1021. 1022. 1023. 1024. 1025. 1026. 1027. 1028. 1029. 1030. 1031. 1032. 1033. 1034. 1035. 1036. 1037. 1038. 1039. 1040. 1041. 1042. 1043. 1044. 1045. 1046. 1047. 1048. 1049. 1050. 1051. 1052. 1053. 1054. 1055. 1056. 1057. 1058. 1059. 1060. 1061. 1062. 1063. 1064. 1065. 1066. 1067. 1068. 1069. 1070. 1071. 1072. 1073. 1074. 1075. 1076. 1077. 1078. 1079. 1080. 1081. 1082. 1083. 1084. 1085. 1086. 1087. 1088. 1089. 1090. 1091. 1092. 1093. 1094. 1095. 1096. 1097. 1098. 1099. 1100. 1101. 1102. 1103. 1104. 1105. 1106. 1107. 1108. 1109. 1110. 1111. 1112. 1113. 1114. 1115. 1116. 1117. 1118. 1119. 1120. 1121. 1122. 1123. 1124. 1125. 1126. 1127. 1128. 1129. 1130. 1131. 1132. 1133. 1134. 1135. 1136. 1137. 1138. 1139. 1140. 1141. 1142. 1143. 1144. 1145. 1146. 1147. 1148. 1149. 1150. 1151. 1152. 1153. 1154. 1155. 1156. 1157. 1158. 1159. 1160. 1161. 1162. 1163. 1164. 1165. 1166. 1167. 1168. 1169. 1170. 1171. 1172. 1173. 1174. 1175. 1176. 1177. 1178. 1179. 1180. 1181. 1182. 1183. 1184. 1185. 1186. 1187. 1188. 1189. 1190. 1191. 1192. 1193. 1194. 1195. 1196. 1197. 1198. 1199. 1200. 1201. 1202. 1203. 1204. 1205. 1206. 1207. 1208. 1209. 1210. 1211. 1212. 1213. 1214. 1215. 1216. 1217. 1218. 1219. 1220. 1221. 1222. 1223. 1224. 1225. 1226. 1227. 1228. 1229. 1230. 1231. 1232. 1233. 1234. 1235. 1236. 1237. 1238. 1239. 1240. 1241. 1242. 1243. 1244. 1245. 1246. 1247. 1248. 1249. 1250. 1251. 1252. 1253. 1254. 1255. 1256. 1257. 1258. 1259. 1260. 1261. 1262. 1263. 1264. 1265. 1266. 1267. 1268. 1269. 1270. 1271. 1272. 1273. 1274. 1275. 1276. 1277. 1278. 1279. 1280. 1281. 1282. 1283. 1284. 1285. 1286. 1287. 1288. 1289. 1290. 1291. 1292. 1293. 1294. 1295. 1296. 1297. 1298. 1299. 1300. 1301. 1302. 1303. 1304. 1305. 1306. 1307. 1308. 1309. 1310. 1311. 1312. 1313. 1314. 1315. 1316. 1317. 1318. 1319. 1320. 1321. 1322. 1323. 1324. 1325. 1326. 1327. 1328. 1329. 1330. 1331. 1332. 1333. 1334. 1335. 1336. 1337. 1338. 1339. 1340. 1341. 1342. 1343. 1344. 1345. 1346. 1347. 1348. 1349. 1350. 1351. 1352. 1353. 1354. 1355. 1356. 1357. 1358. 1359. 1360. 1361. 1362. 1363. 1364. 1365. 1366. 1367. 1368. 1369. 1370. 1371. 1372. 1373. 1374. 1375. 1376. 1377. 1378. 1379. 1380. 1381. 1382. 1383. 1384. 1385. 1386. 1387. 1388. 1389. 1390. 1391. 1392. 1393. 1394. 1395. 1396. 1397. 1398. 1399. 1400. 1401. 1402. 1403. 1404. 1405. 1406. 1407. 1408. 1409. 1410. 1411. 1412. 1413. 1414. 1415. 1416. 1417. 1418. 1419. 1420. 1421. 1422. 1423. 1424. 1425. 1426. 1427. 1428. 1429. 1430. 1431. 1432. 1433. 1434. 1435. 1436. 1437. 1438. 1439. 1440. 1441. 1442. 1443. 1444. 1445. 1446. 1447. 1448. 1449. 1450. 1451. 1452. 1453. 1454. 1455. 1456. 1457. 1458. 1459. 1460. 1461. 1462. 1463. 1464. 1465. 1466. 1467. 1468. 1469. 1470. 1471. 1472. 1473. 1474. 1475. 1476. 1477. 1478. 1479. 1480. 1481. 1482. 1483. 1484. 1485. 1486. 1487. 1488. 1489. 1490. 1491. 1492. 1493. 1494. 1495. 1496. 1497. 1498. 1499. 1500. 1501. 1502. 1503. 1504. 1505. 1506. 1507. 1508. 1509. 1510. 1511. 1512. 1513. 1514. 1515. 1516. 1517. 1518. 1519. 1520. 1521. 1522. 1523. 1524. 1525. 1526. 1527. 1528. 1529. 1530. 1531. 1532. 1533. 1534. 1535. 1536. 1537. 1538. 1539. 1540. 1541. 1542. 1543. 1544. 1545. 1546. 1547. 1548. 1549. 1550. 1551. 1552. 1553. 1554. 1555. 1556. 1557. 1558. 1559. 1560. 1561. 1562. 1563. 1564. 1565. 1566. 1567. 1568. 1569. 1570. 1571. 1572. 1573. 1574. 1575. 1576. 1577. 1578. 1579. 1580. 1581. 1582. 1583. 1584. 1585. 1586. 1587. 1588. 1589. 1590. 1591. 1592. 1593. 1594. 1595. 1596. 1597. 1598. 1599. 1600. 1601. 1602. 1603. 1604. 1605. 1606. 1607. 1608. 1609. 1610. 1611. 1612. 1613. 1614. 1615. 1616. 1617. 1618. 1619. 1620. 1621. 1622. 1623. 1624. 1625. 1626. 1627. 1628. 1629. 1630. 1631. 1632. 1633. 1634. 1635. 1636. 1637. 1638. 1639. 1640. 1641. 1642. 1643. 1644. 1645. 1646. 1647. 1648. 1649. 1650. 1651. 1652. 1653. 1654. 1655. 1656. 1657. 1658. 1659. 1660. 1661. 1662. 1663. 1664. 1665. 1666. 1667. 1668. 1669. 1670. 1671. 1672. 1673. 1674. 1675. 1676. 1677. 1678. 1679. 1680. 1681. 1682. 1683. 1684. 1685. 1686. 1687. 1688. 1689. 1690. 1691. 1692. 1693. 1694. 1695. 1696. 1697. 1698. 1699. 1700. 1701. 1702. 1703. 1704. 1705. 1706. 1707. 1708. 1709. 1710. 1711. 1712. 1713. 1714. 1715. 1716. 1717. 1718. 1719. 1720. 1721. 1722. 1723. 1724. 1725. 1726. 1727. 1728. 1729. 1730. 1731. 1732. 1733. 1734. 1735. 1736. 1737. 1738. 1739. 1740. 1741. 1742. 1743. 1744. 1745. 1746. 1747. 1748. 1749. 1750. 1751. 1752. 1753. 1754. 1755. 1756. 1757. 1758. 1759. 1760. 1761. 1762. 1763. 1764. 1765. 1766. 1767. 1768. 1769. 1770. 1771. 1772. 1773. 1774. 1775. 1776. 1777. 1778. 1779. 1780. 1781. 1782. 1783. 1784. 1785. 1786. 1787. 1788. 1789. 1790. 1791. 1792. 1793. 1794. 1795. 1796. 1797. 1798. 1799. 1800. 1801. 1802. 1803. 1804. 1805. 1806. 1807. 1808. 1809. 1810. 1811. 1812. 1813. 1814. 1815. 1816. 1817. 1818. 1819. 1820. 1821. 1822. 1823. 1824. 1825. 1826. 1827. 1828. 1829. 1830. 1831. 1832. 1833. 1834. 1835. 1836. 1837. 1838. 1839. 1840. 1841. 1842. 1843. 1844. 1845. 1846. 1847. 1848. 1849. 1850. 1851. 1852. 1853. 1854. 1855. 1856. 1857. 1858. 1859. 1860. 1861. 1862. 1863. 1864. 1865. 1866. 1867. 1868. 1869. 1870. 1871. 1872. 1873. 1874. 1875. 1876. 1877. 1878. 1879. 1880. 1881. 1882. 1883. 1884. 1885. 1886. 1887. 1888. 1889. 1890. 1891. 1892. 1893. 1894. 1895. 1896. 1897. 1898. 1899. 1900. 1901. 1902. 1903. 1904. 1905. 1906. 1907. 1908. 1909. 1910. 1911. 1912. 1913. 1914. 1915. 1916. 1917. 1918. 1919. 1920. 1921. 1922. 1923. 1924. 1925. 1926. 1927. 1928. 1929. 1930. 1931. 1932. 1933. 1934. 1935. 1936. 1937. 1938. 1939. 1940. 1941. 1942. 1943. 1944. 1945. 1

思考题 5

1. 答案: 设 X 上的模糊集 \tilde{A} = “严重污染程度”, 则 \tilde{A} 可以表示为:

$$\tilde{A} = 0.3/x_1 + 0.7/x_2 + 0.9/x_3$$

$$\text{或 } \tilde{A} = \{(x_1, 0.3), (x_2, 0.7), (x_3, 0.9)\}$$

2. 答案: 模糊矩阵 $R = (r_{ij})_{n \times n}$, 满足

- 自反性: $r_{ii} = 1$
- 对称性: $r_{ij} = r_{ji}$
- 传递性: $R \cdot R \subseteq R$

$$R \cdot R = \begin{pmatrix} 1.0 & 0.5 & 0.9 \\ 0.9 & 0.5 & 1.0 \\ 0.9 & 0.5 & 1.0 \end{pmatrix} \subseteq \begin{pmatrix} 1.0 & 0.5 & 0.9 \\ 0.9 & 0.5 & 1.0 \\ 0.9 & 0.5 & 1.0 \end{pmatrix} = R$$

则 $R = \begin{pmatrix} 1.0 & 0.5 & 0.9 \\ 0.9 & 0.5 & 1.0 \\ 0.9 & 0.5 & 1.0 \end{pmatrix}$ 是一个模糊等价矩阵。

3. 答案: 模糊矩阵 $R = (r_{ij})_{n \times n}$, 满足

- 自反性: $r_{ii} = 1$
- 对称性: $r_{ij} = r_{ji}$
- 传递性: $R \cdot R \subseteq R$

经过计算 $R^2 = R^3 = R^4$, 故 R^2 就是所求的模糊等价矩阵。

$$R^2 = \begin{pmatrix} 1.0000 & 0.0000 & 0.9412 & 0.9412 \\ 0.0000 & 1.0000 & 0.0000 & 0.0000 \\ 0.9412 & 0.0000 & 1.0000 & 0.9412 \\ 0.9412 & 0.0000 & 0.9412 & 1.0000 \end{pmatrix}$$

4. 答案:

(1) 数据标准化

的变换。根据模糊矩阵的要求将数据压缩到区间 $[0, 1]$ 。

(2) 建立模糊相似矩阵

n), 常使用的方法有 相似系数法、距离法、主观评分法等。

(3) 聚类分析

将 X_1, X_2, \dots, X_n 中的元素 x_i, x_j 按模糊等价关系 R 进行分类。

5. 答案:

(1) 传递闭包法

可以看出共分为类: $\{u_1, u_2, u_3, u_4\}, \{u_5, u_6, u_7\}$ 。

● 当 $\lambda \geq 0.0634$ 时

矩阵的所有元素都变成 1, 只分成 1 类, 是最粗的分类。

(2) 直接聚类法

| | | | | | | | |
|-----|-----|-----|-----|--|-----|-----|-----|
| | 1.0 | 0.8 | 1.0 | | 0.8 | 0.5 | 0.3 |
| | 0.8 | 1.0 | 0.4 | | 0.7 | 0.6 | 0.3 |
| | 1.0 | 0.4 | 1.0 | | 1.0 | 0.6 | 0.5 |
| R | 0.2 | 0.3 | 0.7 | | 0.5 | 0.8 | 0.6 |
| | | | | | 1.0 | 0.2 | 0.7 |
| | 0.5 | 0.6 | 0.6 | | 0.2 | 1.0 | 0.8 |
| | 0.3 | 0.3 | 0.5 | | 0.7 | 0.8 | 1.0 |

R 为论域 U 的模糊相似矩阵

● 由 R 得 U 的模糊等价关系 R^2 为

● 由 R^2 得 U 的模糊等价关系 R^3 为

● 由 R^3 得 U 的模糊等价关系 R^4 为

6. 答案: 利用传递闭包:

| | | | | | | | | | |
|-----|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| | 1.000 0 | 0.932 1 | 0.674 4 | 0.932 1 | 0.674 1 | 0.674 4 | 0.674 4 | 0.932 1 | 0.932 1 |
| | 0.932 1 | 1.000 0 | 0.674 4 | 0.935 9 | 0.674 1 | 0.674 4 | 0.674 4 | 0.965 0 | 0.935 9 |
| | 0.674 4 | 0.674 4 | 1.000 0 | 0.674 4 | 0.795 9 | 0.795 9 | 0.795 9 | 0.674 4 | 0.674 4 |
| | 0.932 1 | 0.935 9 | 0.674 4 | 1.000 0 | 0.674 4 | 0.674 1 | 0.674 4 | 0.935 9 | 0.947 1 |
| R | 0.674 4 | 0.674 4 | 0.795 9 | 0.674 4 | 1.000 0 | 0.990 3 | 0.989 5 | 0.674 4 | 0.674 4 |
| | 0.674 4 | 0.674 4 | 0.795 9 | 0.674 4 | 0.990 3 | 1.000 0 | 0.989 5 | 0.674 4 | 0.674 4 |
| | 0.674 4 | 0.674 4 | 0.795 9 | 0.674 4 | 0.989 5 | 0.989 5 | 1.000 0 | 0.674 4 | 0.674 4 |
| | 0.932 1 | 0.955 0 | 0.674 4 | 0.935 9 | 0.674 4 | 0.674 4 | 0.674 4 | 1.000 0 | 0.935 9 |
| | 0.932 1 | 0.935 9 | 0.674 4 | 0.947 1 | 0.674 4 | 0.674 4 | 0.674 4 | 0.935 9 | 1.000 0 |

● 当 $\lambda \geq 1.0000$ 时

由 R 得 U 的模糊等价关系 R^2 为

■ 当 $\lambda \geq 0.932$ 1 时

| | | |
|-----|---|---|
| | 0 | 0 |
| | 0 | 0 |
| | 1 | 0 |
| | 0 | 0 |
| R | 0 | 1 |
| | 0 | 1 |
| | 0 | 1 |
| | 0 | 0 |
| | 0 | 0 |
| | 0 | 0 |

可以看出共分 3 类: u_1, u_2, u_3 为一类, u_4, u_5, u_6 为一类, u_7, u_8 为一类。

● 当 $\lambda \geq 0.794$ 2 时

0 0 0

$R_1 = 0$

0 0 0 1 1

可以看出共分 2 类: $\{u_1, u_2, u_3, u_4, u_5\}, \{u_6, u_7, u_8, u_9\}$ 。

● 当 $\lambda \geq 0.674$ 4 时

矩阵的所有元素都变成 1, 只分成 1 类, 是最粗的分类。

7. 答案: 利用传递闭包:

| | | | | | | | | | |
|-----|---------|---------|---------|---------|------|---|-----|-----|---|
| R | 0.869 7 | 0.960 1 | 0.950 3 | 0.721 8 | 0.78 | 8 | 809 | 911 | 0 |
| | | | | | | 2 | 186 | 662 | 7 |
| | | | | | | 2 | 616 | 846 | 7 |
| | | | | | | 6 | 772 | 557 | 8 |
| | | | | | | 3 | 798 | 992 | 5 |
| | | | | | | 1 | 186 | 669 | 1 |
| | | | | | | 3 | 346 | 638 | 8 |
| | | | | | | 1 | 548 | 696 | 9 |
| | | | | | | 1 | 840 | 918 | 3 |
| | | | | | | 6 | 825 | 901 | 0 |
| | | | | | | 4 | 743 | 915 | 3 |
| | | | | | | 6 | 913 | 707 | 3 |
| | | | | | | 8 | 925 | 915 | 5 |
| | | | | | | 3 | 841 | 806 | 6 |
| | | | | | | 7 | 479 | 616 | 9 |
| | | | | | | 0 | 570 | 716 | 3 |
| | | | | | | 6 | 000 | 812 | 6 |
| | | | | | | 2 | 960 | 763 | 7 |
| | | | | | | 3 | 812 | 000 | 3 |
| | | | | | | 3 | 695 | 879 | 0 |

从而判别函数:

$$\begin{aligned} W(x) &= (x - \bar{\mu})' \hat{\Sigma}^{-1} (\bar{\mu}_1 - \bar{\mu}_2) \\ &= 308.7812(x_1 - 0.0699) + 3.6016(x_2 - 0.0285) \\ &\quad 136.3149(x_3 - 0.1115) \end{aligned}$$

$$W_1 = 5.5704, W_2 = 12.1639, W_3 = 3.1021,$$

$$W_4 = 17.8223, W = -32.6918$$

G_1 , 样本 2~5 属于 G_2 。

$$\hat{\mu}_1 = \frac{1}{n_1} \sum_{i=1}^5 x_i = \bar{x} = (0.1149 \quad 0.0295 \quad 0.2190)$$

$$\hat{\mu}_2 = \frac{1}{n_2} \sum_{i=6}^8 y_i = \bar{y} = (0.2212 \quad 0.1261 \quad 0.3854)'$$

$$\bar{\mu} = (0.1681 \quad 0.0778 \quad 0.3022)'$$

总体协方差的逆矩阵为:

$$\begin{aligned} \hat{\Sigma} &= \begin{pmatrix} 0.0008 & 0.0002 & -0.0001 \\ 0.0002 & 0.0003 & 0.0001 \\ 0.0004 & 0.0001 & 0.0028 \end{pmatrix} \\ \hat{\Sigma} &= 10^3 \times \begin{pmatrix} 1.5901 & -1.0052 & 0.2360 \\ -1.0052 & 3.9423 & -0.2432 \\ 0.2360 & -0.2432 & 0.3986 \end{pmatrix} \end{aligned}$$

$$\hat{\Sigma}(\bar{\mu}_1 - \bar{\mu}_2) = (-111.3998, -233.1156, -67.9733)'$$

从而判别函数:

$$\begin{aligned} W(x) &= (x - \bar{\mu})' \hat{\Sigma}^{-1} (\bar{\mu}_1 - \bar{\mu}_2) \\ &= -111.3998(x_1 - 0.1681) - 233.1156(x_2 - 0.0778) - \\ &\quad 67.9733(x_3 - 0.3022) \end{aligned}$$

$$W = -29.3751, W_2 = -39.2857, W_3 = 35.5357,$$

$$W_4 = 40.2501, W_5 = 49.2665$$

G_1 , 即样本 1~5 不属于二级。

综合①、②判定样本 1 属于 G_2 , 样本 2~5 属于 G_1 。

(2) Fisher 判别方法

①将一级标准记为 G_1 , 二级标准记为 G_2 .

根据距离判别中的分析数据, 可以得到:

$$y_0 = \frac{1}{2}(\hat{\mu}_1 - \hat{\mu}_2)' \hat{\Sigma}^{-1}(\hat{\mu}_1 + \hat{\mu}_2) = 40.778 0$$

$$y = c'x = (\hat{\mu}_1 - \hat{\mu}_2)' \hat{\Sigma}^{-1}x = 308.781 2x_1 + 3.601 6x_2 - 136.314 9x_3$$

$$\overline{y^{(1)}} = \frac{1}{n_1} \sum_{i=1}^5 y_i^{(1)} = 16.342 8$$

$$\overline{y^{(2)}} = \frac{1}{n_2} \sum_{i=1}^5 y_i^{(2)} = -65.213 2$$

$$\text{即 } \overline{y^{(1)}} > y_0 > \overline{y^{(2)}}$$

将样本 1, 2, 3, 4, 5 的数据分别代入到判别函数中, 得到:

$$y_1 = -46.348 4, y_2 = -28.614 0, y_3 = -37.675 8,$$

$$y_4 = -22.955 5, y_5 = -8.086 2$$

判别结果是一致的.

②将一级标准记为 G_1 , 二级标准记为 G_2 .

根据距离判别中的分析数据, 可以得到:

$$y_0 = \frac{1}{2}(\hat{\mu}_1 - \hat{\mu}_2)' \hat{\Sigma}^{-1}(\hat{\mu}_1 + \hat{\mu}_2) = -57.399 3$$

$$y = c'x = (\hat{\mu}_1 - \hat{\mu}_2)' \hat{\Sigma}^{-1}x = -111.399 8x_1 - 233.115 6x_2 - 67.973 3x_3$$

$$\overline{y^{(1)}} = \frac{1}{n_1} \sum_{i=1}^5 y_i^{(1)} = -34.567 8$$

$$\overline{y^{(2)}} = \frac{1}{n_2} \sum_{i=1}^5 y_i^{(2)} = -80.230 9$$

$$\text{即 } \overline{y^{(1)}} > y_0 > \overline{y^{(2)}}$$

将样本 1, 2, 3, 4, 5 的数据分别代入到判别函数中, 得到:

$$y_1 = 28.024 2, y_2 = -18.113 6, y_3 = 21.863 7,$$

$$y_4 = 17.149 2, y_5 = -8.132 9$$

根据 Fisher 判别准则, 可以判定样本 1~5 属于 G_1 , 即二级.

综合①, ②判定样本 1 属于 G_2 , 样本 2~5 属于 G_1 .

这个结果和距离判别的结果是一致的.

5. 答案:

根据表中的数据, 得到:

$$\hat{\mu}_1 = (4.280 0 \quad 22.140 0 \quad 12.023 3 \quad 51.183 3)'$$

$$\hat{\mu}_2 = (1.070 0 \quad 5.272 5 \quad 2.730 0 \quad 20.675 0)'$$

$$\begin{aligned} \bar{\mu} &= \bar{\mu}_2 = (3.210 \ 0 \quad 15.867 \ 5 \quad 9.293 \ 3 \quad 30.508 \ 3)' \\ \hat{\mu} + \hat{\mu}_2 &= (5.350 \ 0 \quad 27.412 \ 5 \quad 14.753 \ 3 \quad 71.858 \ 3)' \\ \bar{\mu} - (\hat{\mu}_1 - \hat{\mu}_2) &= (2.675 \ 0 \quad 13.706 \ 2 \quad 7.376 \ 7 \quad 35.929 \ 2)' \\ \Sigma &= \begin{pmatrix} 0.772 \ 5 & -0.169 \ 0 & 0.578 \ 1 & 3.178 \ 7 \\ 0.169 \ 0 & 28.661 \ 5 & 13.339 \ 9 & 4.970 \ 6 \\ 0.578 \ 1 & 13.339 \ 9 & 8.418 \ 7 & 20.887 \ 4 \\ 3.178 \ 7 & 4.970 \ 6 & 20.887 \ 4 & 272.165 \ 8 \\ 13.185 \ 0 & 10.817 \ 1 & 22.417 \ 5 & 1.762 \ 1 \\ 10.817 \ 1 & 9.920 \ 8 & -20.497 \ 5 & 1.626 \ 1 \\ -22.417 \ 5 & 20.497 \ 5 & 42.500 \ 2 & -3.370 \ 5 \\ 1.762 \ 1 & 1.626 \ 1 & -3.370 \ 5 & 0.271 \ 2 \end{pmatrix} \\ y_0 &= \frac{1}{2}(\hat{\mu}_1 - \hat{\mu}_2)' \Sigma^{-1}(\hat{\mu}_1 + \hat{\mu}_2) = 460.698 \ 8 \end{aligned}$$

$$= 70.205 \ 1x_1 + 61.183 \ 0x_2 - 125.562 \ 0x_3 + 10.034 \ 8x_4$$

$$y_1^{(1)} = \frac{1}{n_1} \sum_{i=1}^{n_1} y_i^{(1)} = 659.006 \ 8$$

$$y_2^{(2)} = 262.390 \ 9$$

$$\text{即 } y_1^{(1)} > y_0 > y_2^{(2)}$$

将样本 A、B 的数据分别代入到判别函数中, 得到:

$$y_1 = 561.596 \ 4, \quad y_2 = 445.074 \ 8$$

根据 Fisher 判别准则, 样本 A 属于甲地, 样本 B 属于 B 地。

思考题 7

3. 答案:

决策矩阵

$$X = \begin{pmatrix} 2.000 \ 0 & 0.010 \ 0 & 0.200 \ 0 \\ 4.000 \ 0 & 0.025 \ 0 & 0.500 \ 0 \\ 6.000 \ 0 & 0.050 \ 0 & 1.000 \ 0 \\ 10.000 \ 0 & 0.100 \ 0 & 1.500 \ 0 \\ 15.000 \ 0 & 0.200 \ 0 & 2.000 \ 0 \\ 7.100 \ 0 & 0.144 \ 0 & 7.000 \ 0 \\ 5.700 \ 0 & 0.102 \ 0 & 5.270 \ 0 \\ 7.100 \ 0 & 0.107 \ 0 & 3.330 \ 0 \\ 4.000 \ 0 & 0.036 \ 0 & 1.710 \ 0 \\ 4.200 \ 0 & 0.059 \ 0 & 1.900 \ 0 \\ 4.700 \ 0 & 0.078 \ 0 & 2.870 \ 0 \end{pmatrix}$$

五里湖为Ⅲ类。

4. 答案: 原决策矩阵

A

标准化后处 的距离为:

A

0.12 3 0
0.788 6 1
0.593 8 0
0.593 8 0
-0.82. 1 -0
-0.702 0 -1
0.669 3 -1

| | | | | |
|----|--------|-----|---|---|
| 68 | 0.6 0 | 217 | | |
| 52 | 0.6 7 | 271 | - | - |
| 56 | 0.8 9 | 368 | | |
| 4 | 0.6 3 | 659 | - | |
| 71 | -0.2 6 | 213 | | |
| 14 | 0.1 7 | 988 | - | |
| 6 | 0.6 7 | 807 | - | - |
| 42 | 1.5 4 | 281 | | |
| 53 | 2.3 7 | 897 | | |
| 77 | 0.6 1 | 362 | - | - |
| 70 | -0.8 0 | 663 | | |
| 25 | 0.7 3 | 437 | - | |
| 11 | 1.1 7 | 580 | | - |
| 87 | 1.0 4 | 729 | - | - |
| 19 | 1.3 0 | 390 | - | - |
| 74 | 0.7 4 | 341 | | - |
| 13 | 1.3 0 | 929 | - | |
| 6 | 0.5 0 | 379 | | |
| | 0.7 4 | 341 | | |
| 57 | -1.0 9 | 539 | | |
| 52 | -1.0 9 | 266 | - | |
| 86 | -0.8 7 | 086 | - | |

计算标准化数据矩阵 A 的协方差矩阵

$$C = \begin{pmatrix} 1.000 & 0 & 0.525 & 6 & 0.690 & 0 & 0.520 & 7 & 0.535 & 1 & 0.268 & 0 & 0.061 & 1 \\ 0.525 & 6 & 1.000 & 0 & 0.652 & 0 & 0.871 & 6 & 0.892 & 0 & 0.494 & 9 & 0.056 & 5 \\ 0.690 & 0 & 0.652 & 0 & 1.000 & 0 & 0.838 & 5 & 0.593 & 9 & 0.695 & 3 & 0.058 & 3 \\ 0.520 & 7 & 0.871 & 6 & 0.838 & 5 & 1.000 & 0 & 0.785 & 2 & 0.655 & 3 & 0.023 & 4 \\ 0.535 & 1 & 0.892 & 0 & 0.593 & 9 & 0.785 & 2 & 1.000 & 0 & 0.505 & 1 & 0.029 & 7 \\ 0.268 & 0 & 0.494 & 9 & 0.695 & 3 & 0.655 & 3 & 0.505 & 1 & 1.000 & 0 & 0.054 & 6 \\ 0.061 & 1 & 0.056 & 5 & 0.058 & 3 & 0.023 & 4 & 0.029 & 7 & 0.054 & 6 & 1.000 & 0 \end{pmatrix}$$

求 C 的特征值 $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \lambda_4 \geq \lambda_5 \geq \lambda_6 \geq \lambda_7$, 以及对应的特征向量 $u_1, u_2, u_3, u_4, u_5, u_6, u_7$, 要求它们是标准正交的。

$$U = \begin{pmatrix} 0.137 & 9 & -0.122 & 4 & 0.360 & 7 & 0.180 & 1 & 0.135 & 1 & -0.111 & 4 & 0.112 & 5 \\ 0.178 & 8 & -0.067 & 3 & 0.018 & 5 & -0.195 & 0 & -0.053 & 2 & 0.280 & 2 & 0.226 & 1 \\ 0.177 & 2 & 0.046 & 7 & 0.003 & 0 & 0.195 & 0 & -0.169 & 4 & 0.186 & 5 & 0.286 & 6 \\ 0.187 & 8 & 0.020 & 2 & -0.066 & 4 & -0.045 & 9 & -0.230 & 4 & 0.002 & 1 & 0.351 & 8 \\ 0.173 & 1 & 0.016 & 1 & 0.048 & 6 & 0.208 & 2 & 0.199 & 7 & 0.273 & 7 & 0.006 & 5 \\ 0.112 & 4 & 0.125 & 9 & -0.350 & 6 & 0.141 & 8 & 0.207 & 9 & 0.107 & 1 & 0.035 & 1 \\ 0.002 & 8 & 0.601 & 4 & 0.152 & 2 & 0.034 & 0 & 0.004 & 2 & 0.038 & 9 & 0.001 & 8 \end{pmatrix}$$

$$= (u_1, u_2, u_3, u_4, u_5, u_6, u_7)$$

$$\lambda_1 = 4.224 \quad \lambda_2 = 1.034 \quad \lambda_3 = 0.722 \quad \lambda_4 = 0.619 \quad \lambda_5 = 0.240 \quad \lambda_6 = 0.078 \quad \lambda_7 = 0.050$$

累计贡献率

各个成分的累计贡献率分别为:

$$0.603 \quad 0.751 \quad 0.854 \quad 0.947 \quad 0.981 \quad 0.992 \quad 1.000$$

求第一主成分 F , 有

$$F = Au_1 =$$

$$\begin{pmatrix} 1.399 & 7 & 0.740 & 2 & 0.770 & 9 & 0.088 & 8 & -0.116 & 3 & -0.038 & 0 & 0.334 & 4 & 1.301 & 4 \\ 1.488 & 5 & 0.012 & 2 & -0.653 & 7 & 0.329 & 9 & 0.859 & 0 & 0.839 & 8 & 1.049 & 5 & 0.432 & 2 \\ 1.041 & 7 & -0.403 & 7 & -0.678 & 7 & 0.974 & 3 & 0.978 & 2 & 0.929 & 0 \end{pmatrix}$$

$$F = \begin{pmatrix} 1.399 & 7 & 0.740 & 2 & 0.770 & 9 & 0.088 & 8 & 0.116 & 3 & 0.038 & 0 & 0.334 & 4 & 1.301 & 4 \\ 1.488 & 5 & 0.042 & 2 & -0.653 & 7 & 0.329 & 9 & 0.859 & 0 & 0.839 & 8 & 1.049 & 5 & 0.432 & 2 \\ 1.041 & 7 & -0.403 & 7 & 0.678 & 7 & 0.974 & 3 & -0.978 & 2 & 0.929 & 0 \end{pmatrix}$$

按 F , 由大到小的顺序排列方案的优先次序, 结果是:

$$e_1 \succ e_{11} \succ e_9 \succ e_{10} \succ e_{15} \succ e_{12} \succ e_{14} \succ e_{16} \succ e_7 \succ e_5 \succ e_{13} \succ e_2 \succ e_3 \succ e_4 \succ e_6 \succ e_8$$

下, 各个采样点由优到劣的排列顺序

根据式(8.17)

$$A = (\sqrt{\lambda_1} e_1, \sqrt{\lambda_2} e_2, \sqrt{\lambda_3} e_3) = \begin{bmatrix} 0.632\ 5 & 0.471\ 9 & 0.336\ 3 \\ 0.688\ 1 & -0.264\ 2 & 0.594\ 5 \\ -0.846\ 6 & -0.081\ 4 & 0.135\ 1 \\ 0.737\ 5 & -0.389\ 4 & 0.442\ 5 \\ 0.109\ 3 & 0.929\ 5 & 0.142\ 5 \\ 0.735\ 8 \\ 0.896\ 7 \end{bmatrix}$$

其同度,

$$h = \begin{bmatrix} 0.741\ 5 \\ 0.891\ 3 \\ 0.896\ 3 \end{bmatrix}$$

各个公因子 f_j 对所有变量的贡献 $g = (2.145\ 9 \quad 1.314\ 7 \quad 0.700\ 9)$

$$A_1^* = \begin{bmatrix} 0.090\ 1 & 0.546\ 3 & -0.655\ 2 \\ 0.937\ 7 & -0.079\ 1 & -0.105\ 9 \\ 0.646\ 9 & 0.246\ 1 & 0.512\ 4 \\ 0.248\ 5 & 0.287\ 5 & 0.864\ 2 \\ 0.002\ 3 & 0.943\ 0 & 0.084\ 2 \end{bmatrix}$$

各个公因子 f_j 对所有变量的贡献 $g = (1.367\ 5 \quad 1.337\ 1 \quad 1.457\ 0)$

$$0.735\ 8$$

$$0.896\ 7$$

其同度不变,

$$h = \begin{bmatrix} 0.741\ 5 \\ 0.891\ 3 \\ 0.896\ 3 \end{bmatrix}$$

第六步,求因子得分。

(1)求特殊向量方差 ψ 。

$$\psi = \begin{bmatrix} 0.264\ 2 & 0.000\ 0 & 0.000\ 0 & 0.000\ 0 & 0.000\ 0 \\ 0.000\ 0 & 0.103\ 3 & 0.000\ 0 & 0.000\ 0 & 0.000\ 0 \\ 0.000\ 0 & 0.000\ 0 & 0.258\ 5 & 0.000\ 0 & 0.000\ 0 \\ 0.000\ 0 & 0.000\ 0 & 0.000\ 0 & 0.108\ 7 & 0.000\ 0 \\ 0.000\ 0 & 0.000\ 0 & 0.000\ 0 & 0.000\ 0 & 0.103\ 7 \end{bmatrix}$$

(2)运用式(8.30),得到因子得分 F ,结果见下表。

因子得分表

| 样本序号 | 因子得分 | | |
|------|----------|----------|----------|
| | f_1 | f_2 | f_3 |
| 1 | -0.990 1 | 0.618 0 | 0.600 8 |
| 2 | 1.524 0 | -0.750 5 | 1.263 1 |
| 3 | 0.052 1 | 0.950 2 | 0.670 6 |
| 4 | -1.108 7 | 1.292 1 | -1.844 4 |
| 5 | 0.472 1 | -0.739 0 | 0.066 3 |
| 6 | -0.122 6 | -1.665 0 | -1.662 4 |
| 7 | 0.046 5 | 1.058 2 | 0.485 8 |
| 8 | 1.206 0 | 0.985 4 | 0.278 6 |
| 9 | 1.598 9 | 0.100 4 | -0.153 1 |
| 10 | 0.462 9 | 0.251 4 | 0.294 6 |

4. 答案:

2. 求样本的协方差矩阵 S 和样本的均值向量 \bar{X} 。

| | | | | | | |
|-----|----------|----------|----------|----------|---------|----------|
| S | 0.045 1 | 0.033 2 | 0.979 7 | -1.290 7 | 767 1 | -0.367 0 |
| | -0.376 1 | 1.316 3 | 1.292 5 | -0.557 5 | 0.548 5 | -0.397 7 |
| | 1.037 9 | 2.002 2 | 0.601 0 | 0.053 5 | 0.017 5 | -0.323 1 |
| | 1.278 6 | 0.453 5 | -0.090 6 | -0.048 4 | 250 1 | -0.352 3 |
| | 1.729 9 | -0.829 6 | 1.045 5 | 1.580 9 | 705 7 | 0.327 2 |
| | 0.526 5 | -0.564 1 | 1.292 5 | 0.460 8 | 454 2 | 2.474 1 |
| | -0.436 2 | 0.188 0 | 0.041 2 | 0.970 0 | 297 0 | -0.337 7 |
| | 0.827 3 | 0.0996 | -1.111 4 | -1.168 5 | 504 9 | 0.369 1 |

第二步,求样本的相关矩阵 R 。

| | | | | | | |
|-----|----------|----------|----------|----------|---------|---------|
| R | 1.000 0 | -0.556 6 | -0.443 4 | 0.249 3 | .519 5 | .213 9 |
| | -0.556 6 | 1.000 0 | -0.067 3 | 0.091 9 | .377 0 | .213 6 |
| | 0.443 4 | -0.067 3 | 1.000 0 | 0.123 1 | .081 9 | .517 1 |
| | 0.249 3 | -0.091 9 | 0.123 1 | 1.000 0 | .145 0 | .203 1 |
| | 0.519 5 | 0.377 0 | 0.081 9 | -0.145 0 | 1.000 0 | .182 5 |
| | 0.213 9 | -0.213 6 | 0.517 1 | 0.203 1 | .182 5 | 1.000 0 |

第三步,求 R 的特征值 λ 及其相应的特征向量。

R 的特征值及其累计方差贡献率分别为:

| R 的特征值及其累计方差贡献率 | | | | | | |
|-----------------|---------|---------|---------|---------|---------|---------|
| 特征值 | 2.182 8 | 1.606 7 | 0.919 5 | 0.646 8 | 0.568 9 | 0.075 2 |
| 累计方差贡献率/% | 0.363 8 | 0.631 6 | 0.784 8 | 0.892 6 | 0.987 5 | 1.000 0 |

所以, 取前三个主成分, 即 F_1, F_2, F_3 即可代表全部信息了。

它们对应的特征向量:

$$\begin{aligned} e_1 &= (0.563\ 1 \quad -0.504\ 4 \quad 0.023\ 7 \quad 0.277\ 1 \quad -0.494\ 1 \quad 0.327\ 1)' \\ e_2 &= (0.355\ 6 \quad 0.048\ 6 \quad 0.743\ 0 \quad 0.169\ 6 \quad 0.032\ 0 \quad 0.537\ 9)' \\ e_3 &= (0.072\ 3 \quad 0.343\ 6 \quad -0.164\ 7 \quad 0.899\ 6 \quad 0.194\ 4 \quad -0.051\ 0)' \end{aligned}$$

第四步, 求因子载荷矩阵 A 。

根据式(8.17)

$$A = (\sqrt{\lambda_1} e_1, \sqrt{\lambda_2} e_2, \sqrt{\lambda_3} e_3) = \begin{vmatrix} 0.831\ 9 & -0.450\ 8 & 0.069\ 3 \\ 0.745\ 3 & 0.061\ 6 & 0.329\ 5 \\ 0.035\ 0 & 0.941\ 8 & -0.157\ 9 \\ 0.409\ 4 & 0.215\ 0 & 0.862\ 6 \\ -0.730\ 1 & 0.040\ 6 & 0.186\ 4 \\ 0.483\ 3 & 0.681\ 8 & 0.048\ 9 \\ 0.900\ 0 \\ 0.667\ 8 \\ 0.913\ 2 \\ 0.957\ 9 \\ 0.569\ 4 \\ 0.700\ 8 \end{vmatrix}$$

共同度,

$h^2 =$

各个公因子 f_i 对所有变量的贡献 g_i (2.182 8 1.606 7 0.919 5)

所以, 取前三个主成分, 即 F_1, F_2, F_3 即可代表全部信息了。

$$A_2^* = \begin{vmatrix} 0.856\ 5 & -0.293\ 9 & 0.283\ 1 \\ -0.801\ 6 & -0.142\ 3 & 0.070\ 0 \\ 0.144\ 2 & 0.944\ 5 & 0.018\ 7 \\ 0.073\ 7 & 0.115\ 7 & 0.969\ 1 \\ -0.740\ 4 & -0.132\ 0 & -0.061\ 4 \\ 0.300\ 2 & 0.755\ 7 & 0.198\ 8 \end{vmatrix}$$

各个公因子 f_i 对所有变量的贡献 g_i (2.040 6 1.600 6 1.067 8)

第六步, 求因子得分。

(1) 求特殊向量方差 ψ 。

$$\psi = \begin{pmatrix} 0.1000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.3322 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0868 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0421 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.4306 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.2992 \end{pmatrix}$$

(2) 运用式(8.30), 得到因子得分 F , 结果见下表。

因子得分表

| 因子 | 因子得分 | | |
|----|---------|--------|--------|
| | f_1 | f_2 | f_3 |
| 1 | -0.0217 | 0.0074 | 0.0643 |
| 2 | -0.0338 | 0.0440 | 0.1198 |
| 3 | 0.0862 | 0.0391 | 0.1998 |
| 4 | -0.1248 | 0.0136 | 0.1940 |
| 5 | -0.0399 | 0.0052 | 0.3363 |
| 6 | 0.1139 | 0.6685 | 0.3269 |
| 7 | -0.0688 | 0.0274 | 0.2774 |
| 8 | -0.0497 | 0.0050 | 0.0854 |

5. 答案:

第一步, 计算 $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8$ 的均值, 如表 8.3.1 所示。

| | | | | | | | |
|-------------|---------|---------|---------|---------|---------|---------|---------|
| $\bar{X} =$ | 1.1682 | 1.7279 | 0.5658 | 0.5287 | 0.5866 | 2.3020 | -0.4214 |
| | 1.3811 | 1.3663 | 0.7435 | -0.9154 | -1.0987 | -0.1703 | -0.7568 |
| | 0.5655 | -0.5754 | -0.5777 | 0.2009 | 0.1282 | -0.6792 | -0.5802 |
| | 0.0023 | 0.0601 | -0.7435 | -0.5830 | -0.6020 | -0.5285 | -0.7547 |
| | 0.9913 | -0.8837 | 1.6852 | 1.6497 | 1.6352 | 0.5825 | 1.5833 |
| | 0.9812 | 0.8867 | 0.7273 | 0.8511 | 0.7292 | 0.5012 | 0.6415 |
| | -1.0522 | 0.9756 | 1.4517 | 1.1668 | 1.2352 | -0.5166 | 1.5871 |
| | 1.1581 | 0.4621 | 0.5753 | 0.9196 | -0.8316 | 0.4465 | -0.5514 |
| | -0.1194 | -0.2849 | 0.6584 | 0.9217 | -0.6688 | -0.9351 | -0.7474 |

第二步, 求样本的相关矩阵 R 。

| | | | | | | | |
|-------|---------|--------|---------|---------|---------|---------|---------|
| $R =$ | 1.0000 | 0.9466 | -0.7328 | -0.8312 | 0.8569 | 0.4904 | 0.7042 |
| | 0.9466 | 1.0000 | -0.6659 | -0.7278 | 0.7703 | 0.6270 | -0.6285 |
| | 0.7328 | 0.6659 | 1.0000 | 0.9407 | 0.9526 | 0.0206 | 0.9960 |
| | -0.8312 | 0.7278 | 0.9407 | 1.0000 | 0.9915 | 0.0633 | 0.9324 |
| | 0.8569 | 0.7703 | 0.9526 | 0.9915 | 1.0000 | -0.0868 | 0.9431 |
| | 0.4904 | 0.6270 | -0.0206 | 0.0633 | -0.0868 | 1.0000 | 0.0247 |
| | -0.7042 | 0.6285 | 0.9960 | 0.9324 | 0.9431 | 0.0247 | 1.0000 |
| | | | | | | | |

第二步,求 R 的特征值 λ 及其相应的特征向量。

R 的特征值及其累计方差贡献率分别为:

| R 的特征值及其累计方差贡献率 | | | | | | | |
|-----------------|---------|---------|---------|---------|---------|---------|---------|
| 特征值 | 5.276 9 | 1.446 3 | 0.187 0 | 0.058 5 | 0.026 5 | 0.002 6 | 0.002 1 |
| 累计方差贡献率 % | 0.733 8 | 0.960 5 | 0.987 2 | 0.995 5 | 0.999 3 | 0.999 7 | 1.000 0 |

以了。

它们对应的特征向量:

$$e_1 = (0.399\ 5 \quad 0.376\ 1 \quad 0.408\ 3 \quad 0.419\ 5 \quad 0.426\ 8 \quad -0.115\ 5 \quad 0.401\ 0)'$$

$$e_2 = (0.260\ 1 \quad 0.384\ 1 \quad 0.236\ 3 \quad 0.168\ 2 \quad 0.143\ 5 \quad 0.777\ 8 \quad 0.274\ 0)'$$

第四步,求因子载荷矩阵 A 。

根据式(8.17)

$$A = (\sqrt{\lambda_1} e_1 \quad \sqrt{\lambda_2} e_2) = \begin{pmatrix} -0.917\ 7 & 0.312\ 9 \\ -0.863\ 9 & 0.461\ 9 \\ 0.937\ 9 & 0.281\ 2 \\ 0.963\ 7 & 0.202\ 3 \\ 0.980\ 3 & 0.172\ 6 \\ -0.265\ 2 & 0.935\ 1 \\ 0.921\ 2 & 0.329\ 1 \\ 0.940\ 1 \\ 0.959\ 7 \\ 0.960\ 1 \end{pmatrix}$$

共同度:

$$h^2 = \begin{pmatrix} 0.969\ 6 \\ 0.990\ 8 \\ 0.945\ 4 \\ 0.957\ 2 \end{pmatrix}$$

各个公因子 f_i 对所有变量的贡献 g_i (5.276 9 1.446 3)

$$A_i' = \begin{pmatrix} 0.733\ 4 & 0.631\ 9 \\ 0.630\ 0 & 0.750\ 2 \\ 0.976\ 3 & -0.087\ 2 \\ 0.969\ 8 & -0.170\ 8 \\ 0.971\ 2 & 0.204\ 5 \\ 0.101\ 9 & 0.967\ 0 \\ 0.977\ 7 & 0.037\ 0 \end{pmatrix}$$

各个公因子 f_i 对所有变量的贡献 $g = (4.746\ 5\ 1.976\ 7)$

第六步,求因子得分。

(1)求特殊向量方差 ψ 。

$$\psi = \begin{bmatrix} 0.059\ 9 & 0.000\ 0 & 0.000\ 0 & 0.000\ 0 & 0.000\ 0 & 0.000\ 0 & 0.000\ 0 \\ 0.000\ 0 & 0.040\ 3 & 0.000\ 0 & 0.000\ 0 & 0.000\ 0 & 0.000\ 0 & 0.000\ 0 \\ 0.000\ 0 & 0.000\ 0 & 0.039\ 6 & 0.000\ 0 & 0.000\ 0 & 0.000\ 0 & 0.000\ 0 \\ 0.000\ 0 & 0.000\ 0 & 0.000\ 0 & 0.030\ 4 & 0.000\ 0 & 0.000\ 0 & 0.000\ 0 \\ 0.000\ 0 & 0.000\ 0 & 0.000\ 0 & 0.000\ 0 & 0.009\ 2 & 0.000\ 0 & 0.000\ 0 \\ 0.000\ 0 & 0.000\ 0 & 0.000\ 0 & 0.000\ 0 & 0.000\ 0 & 0.054\ 6 & 0.000\ 0 \\ 0.000\ 0 & 0.000\ 0 & 0.000\ 0 & 0.000\ 0 & 0.000\ 0 & 0.000\ 0 & 0.042\ 8 \end{bmatrix}$$

(2)运用式(8.30),得到因子得分 F ,结果见下表。

因子得分表

| 样本序号 | 因子得分 | |
|------|---------|---------|
| | f_1 | f_2 |
| 1 | 1.208 6 | 0.839 4 |
| 2 | 0.755 4 | 0.545 3 |
| 3 | 1.658 0 | 0.678 1 |
| 4 | 1.114 3 | 0.584 6 |
| 5 | 2.894 9 | 1.033 8 |
| 6 | 2.131 3 | 0.791 2 |
| 7 | 2.519 4 | 0.869 4 |
| 8 | 0.971 8 | 0.639 0 |
| 9 | 1.051 0 | 0.538 6 |
| 10 | 1.208 6 | 0.839 4 |

思考题 10

设 $X = (X_1, X_2, \dots, X_n)$ 是 n 个随机变量, $Y = (Y_1, Y_2, \dots, Y_m)$ 是 m 个随机变量, X 和 Y 的联合分布函数为 $F(x, y)$, $F(x, y)$ 在 $(-\infty, +\infty) \times (-\infty, +\infty)$ 上满足下列条件:

(1) $F(x, y)$ 在 $(-\infty, +\infty) \times (-\infty, +\infty)$ 上非负, 且 $F(x, y) \leq 1$;

$$\text{Cov}(x, y) = E[(x - E(x))(y - E(y))]$$

试证明: $\text{Cov}(x, y) = E[(x - E(x))(y - E(y))]$ 是 x 和 y 的协方差函数, 且 $\text{Cov}(x, y) = E[(x - E(x))(y - E(y))]$ 是 x 和 y 的协方差函数。

$$\begin{aligned}
 y(3) = & \frac{1}{2 \times 20} [(15-15)^2 + (18-10)^2 + (13-17)^2 + (15-16)^2 + (10-18)^2 + \\
 & (11-21)^2 + (12-18)^2 + (14-16)^2 + (17-21)^2 + (19-18)^2 + (15-12)^2 + \\
 & (13-17)^2 + (18-14)^2 + (15-19)^2 + (16-15)^2 + (20-23)^2 + (15-18)^2 + \\
 & (17-21)^2 + (10-16)^2 + (16-18)^2] = \frac{434}{40} = 10.85
 \end{aligned}$$

$$\begin{aligned}
 y(4) = & \frac{1}{2 \times 10} [(15-10)^2 + (13-16)^2 + (10-21)^2 + (12-16)^2 + (17-18)^2 + \\
 & (15-17)^2 + (18-19)^2 + (16-23)^2 + (15-21)^2 + (10-18)^2]
 \end{aligned}$$

4. 答案:

根据公式(10-32), 可以计算:

$$y = \frac{36 \times 5.89 + 26 \times 11.88 + 20 \times 10.85 + 10 \times 16.30}{36 + 26 + 20 + 10} = 9.79$$

$$\bar{x} = \frac{36 \times 2 + 26 \times 4 + 20 \times 6 + 10 \times 8}{36 + 26 + 20 + 10} = 4.09$$

$$x = \frac{36 \times 2 + 26 \times 4 + 20 \times 6 + 10 \times 8}{36 + 26 + 20 + 10} = 123.83$$

$$10 \times (8 - 4.09)^2 = 383.30$$

$$\begin{aligned}
 L_{20} = & 36 \times (8 - 123.83)^2 + 26 \times (64 - 123.83)^2 + 20 \times (216 - 123.83)^2 + \\
 & 10 \times (512 - 123.83)^2 = 2\,252\,733.22
 \end{aligned}$$

$$\begin{aligned}
 L_{10} = & L_{20} - 36 \times (2 - 4.09) \times (8 - 123.83) + 26 \times (4 - 4.09) \times (64 - 123.83) + 20 \times (6 - 4.09) \times \\
 & (216 - 123.83) + 10 \times (8 - 4.09) \times (512 - 123.83) = 27\,553.39
 \end{aligned}$$

$$\begin{aligned}
 L_{11} = & 36 \times (2 - 4.09) \times (5.89 - 9.79) + 26 \times (4 - 4.09) \times (11.88 - 9.79) + 20 \times (6 - 4.09) \times \\
 & (10.85 - 9.79) + 10 \times (8 - 4.09) \times (16.30 - 9.79) = 583.56
 \end{aligned}$$

$$\begin{aligned}
 L_{21} = & 36 \times (8 - 123.83) \times (5.89 - 9.79) + 26 \times (64 - 123.83) \times (11.88 - 9.79) + 20 \times (216 - \\
 & 123.83) \times (10.85 - 9.79) + 10 \times (512 - 123.83) \times (16.30 - 9.79) = 40\,235.24
 \end{aligned}$$

$$\begin{aligned}
 L_{22} = & 36 \times (5.89 - 9.79)^2 + 26 \times (11.88 - 9.79)^2 + 20 \times (10.85 - 9.79)^2 + 10 \times (16.30 - 9.79)^2 \\
 & = 1\,107.40
 \end{aligned}$$

从而可求得,

$$b_2 = \frac{583.56 \times 2\,252\,733.22 - 40\,235.24 \times 27\,553.39}{383.30 \times 2\,252\,733.22 - 27\,553.39 \times 27\,553.39} = 1.98$$

$$b_1 = \frac{40\,235.24 \times 383.30 - 583.56 \times 27\,553.39}{383.30 \times 2\,252\,733.22 - 27\,553.39 \times 27\,553.39} = 0.01$$

$$b_0 = 9.79 - 1.98 \times 4.09 + 0.01 \times 123.83 = 2.93$$

由于 $b_0 > 0$, $b_1 > 0$, $b_2 < 0$, 此时球状模型中三个参数 C_0 , C 和 a 分别为:

因此, 球状模型为:

$$\gamma(h) = \begin{cases} 0 & (h=0) \\ 2.93 + 11.10 \times \left(\frac{3}{2} \times \frac{h}{8.41} - \frac{1}{2} \times \frac{h^3}{8.41^3} \right) & (0 < h < 8.41) \\ 14.03 & (h > 8.41) \end{cases}$$

2. 答案:

$$c^*(h) = \begin{cases} 14.03 & (h=0) \\ 11.10 \times \left[1 - \left(\frac{3}{2} \times \frac{h}{8.41} - \frac{1}{2} \times \frac{h^3}{8.41^3} \right) \right] & (0 < h < 8.41) \\ 0 & (h > 8.41) \end{cases}$$

因此,

$$c_1 = c_{02} = c_{03} = c_{04} = 14.03$$

$$= 3.11$$

$$c_{14} = c_{04} = c_{02} = 14.03 - \gamma(\sqrt{4^2}) = 14.03 - \gamma(4)$$

$$= 14.03 - \left[2.93 + 11.10 \times \left(\frac{3}{2} \times \right. \right.$$

$$c_{01} = 14.03 - \gamma(\sqrt{2^2}) = 14.03 - \gamma(2)$$

$$= 14.03 - \left[2.93 + 11.10 \times \left(\frac{3}{2} \times \frac{2}{8.41} - \frac{1}{2} \times \frac{2^3}{8.41^3} \right) \right] = 7.21$$

$$c_{21} = c_{02} = c_{03} = 14.03 - \gamma(\sqrt{2^2 + 2^2}) = 14.03 - \gamma(2\sqrt{2})$$

$$= 14.03 - \left[2.93 + 11.10 \times \left(\frac{3}{2} \times \frac{2\sqrt{2}}{8.41} - \frac{1}{2} \times \frac{(2\sqrt{2})^3}{8.41^3} \right) \right] = 5.66$$

$$c_{05} = c_{04} = 14.03 - \gamma(\sqrt{8^2 + 2^2}) = 14.03 - \gamma(2\sqrt{17})$$

$$= 14.03 - \left[2.93 + 11.10 \times \left(\frac{3}{2} \times \frac{2\sqrt{17}}{8.41} - \frac{1}{2} \times \frac{(2\sqrt{17})^3}{8.41^3} \right) \right] = 0.00$$

$$c_{04} = c_{03} = 14.03 - \gamma(\sqrt{6^2 + 4^2}) = 14.03 - \gamma(2\sqrt{13})$$

$$= 14.03 - \left[2.93 + 11.10 \times \left(\frac{3}{2} \times \frac{2\sqrt{13}}{8.41} - \frac{1}{2} \times \frac{(2\sqrt{13})^3}{8.41^3} \right) \right] = 0.33$$

可以得到克立格方程组:

$$\begin{array}{l} \lambda_1 \left| \begin{array}{ccccc} 14.03 & 3.11 & 3.11 & 3.77 & 1.00 \end{array} \right| \begin{array}{l} 1 \\ 17.21 \end{array} \left| \begin{array}{c} 0.40 \\ 0.06 \end{array} \right. \\ \left. \begin{array}{ccccc} 3.11 & 14.03 & 5.66 & 0.00 & 1.00 \end{array} \right| \begin{array}{c} 3.77 \\ 5.66 \end{array} \left| \begin{array}{c} 0.40 \\ 0.28 \end{array} \right. \\ \lambda_2 \left| \begin{array}{ccccc} 3.11 & 5.66 & 14.03 & 0.33 & 1.00 \end{array} \right| \begin{array}{c} 5.66 \\ 3.11 \end{array} \left| \begin{array}{c} 0.28 \\ 0.09 \end{array} \right. \\ \lambda_3 \left| \begin{array}{ccccc} 3.77 & 0.00 & 0.33 & 14.03 & 1.00 \end{array} \right| \begin{array}{c} 3.11 \\ 1.00 \end{array} \left| \begin{array}{c} 0.09 \\ 0.17 \end{array} \right. \\ \mu \left| \begin{array}{ccccc} 1.00 & 1.00 & 1.00 & 1.00 & 0.00 \end{array} \right| \begin{array}{c} 1.00 \\ 0.17 \end{array} \left| \begin{array}{c} 0.17 \\ 0.17 \end{array} \right. \end{array}$$

附录

附表1 标准正态分布表



$$\frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = P(Z \leq z_p) = \Phi$$

| z | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 |
|-----|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 0.0 | 0.5000 | 0.5039 | 0.5078 | 0.5117 | 0.5156 | 0.5195 | 0.5234 | 0.5273 | 0.5311 | 0.5350 |
| 0.1 | 0.5398 | 0.5437 | 0.5476 | 0.5515 | 0.5554 | 0.5593 | 0.5632 | 0.5671 | 0.5710 | 0.5748 |
| 0.2 | 0.5787 | 0.5825 | 0.5863 | 0.5901 | 0.5938 | 0.5976 | 0.6013 | 0.6051 | 0.6088 | 0.6125 |
| 0.3 | 0.6163 | 0.6201 | 0.6238 | 0.6275 | 0.6311 | 0.6348 | 0.6385 | 0.6422 | 0.6458 | 0.6495 |
| 0.4 | 0.6532 | 0.6568 | 0.6604 | 0.6641 | 0.6677 | 0.6712 | 0.6748 | 0.6783 | 0.6819 | 0.6854 |
| 0.5 | 0.6888 | 0.6924 | 0.6959 | 0.6994 | 0.7029 | 0.7064 | 0.7098 | 0.7133 | 0.7167 | 0.7201 |
| 0.6 | 0.7234 | 0.7268 | 0.7301 | 0.7334 | 0.7367 | 0.7400 | 0.7433 | 0.7465 | 0.7497 | 0.7529 |
| 0.7 | 0.7560 | 0.7591 | 0.7623 | 0.7654 | 0.7685 | 0.7715 | 0.7745 | 0.7774 | 0.7804 | 0.7833 |
| 0.8 | 0.7863 | 0.7892 | 0.7921 | 0.7950 | 0.7978 | 0.8006 | 0.8034 | 0.8062 | 0.8089 | 0.8116 |
| 0.9 | 0.8143 | 0.8169 | 0.8195 | 0.8221 | 0.8246 | 0.8271 | 0.8296 | 0.8320 | 0.8345 | 0.8370 |
| 1.0 | 0.8394 | 0.8418 | 0.8441 | 0.8465 | 0.8488 | 0.8511 | 0.8534 | 0.8557 | 0.8579 | 0.8601 |
| 1.1 | 0.8623 | 0.8645 | 0.8667 | 0.8688 | 0.8709 | 0.8729 | 0.8749 | 0.8769 | 0.8788 | 0.8808 |
| 1.2 | 0.8828 | 0.8847 | 0.8866 | 0.8885 | 0.8903 | 0.8921 | 0.8938 | 0.8956 | 0.8973 | 0.8990 |
| 1.3 | 0.9008 | 0.9024 | 0.9041 | 0.9058 | 0.9074 | 0.9089 | 0.9106 | 0.9122 | 0.9137 | 0.9152 |
| 1.4 | 0.9167 | 0.9182 | 0.9197 | 0.9212 | 0.9226 | 0.9241 | 0.9255 | 0.9269 | 0.9282 | 0.9296 |
| 1.5 | 0.9309 | 0.9323 | 0.9336 | 0.9349 | 0.9362 | 0.9375 | 0.9388 | 0.9401 | 0.9413 | 0.9426 |
| 1.6 | 0.9438 | 0.9450 | 0.9461 | 0.9473 | 0.9484 | 0.9495 | 0.9506 | 0.9517 | 0.9527 | 0.9537 |
| 1.7 | 0.9547 | 0.9557 | 0.9566 | 0.9576 | 0.9585 | 0.9594 | 0.9603 | 0.9611 | 0.9619 | 0.9628 |
| 1.8 | 0.9636 | 0.9644 | 0.9651 | 0.9658 | 0.9665 | 0.9672 | 0.9679 | 0.9686 | 0.9692 | 0.9699 |
| 1.9 | 0.9706 | 0.9713 | 0.9719 | 0.9726 | 0.9732 | 0.9738 | 0.9744 | 0.9750 | 0.9756 | 0.9761 |
| 2.0 | 0.9766 | 0.9772 | 0.9778 | 0.9783 | 0.9788 | 0.9793 | 0.9798 | 0.9803 | 0.9808 | 0.9812 |
| 2.1 | 0.9817 | 0.9821 | 0.9826 | 0.9830 | 0.9834 | 0.9838 | 0.9842 | 0.9846 | 0.9850 | 0.9854 |
| 2.2 | 0.9858 | 0.9861 | 0.9864 | 0.9868 | 0.9871 | 0.9874 | 0.9877 | 0.9880 | 0.9883 | 0.9886 |
| 2.3 | 0.9889 | 0.9891 | 0.9894 | 0.9896 | 0.9898 | 0.9900 | 0.9902 | 0.9904 | 0.9906 | 0.9908 |
| 2.4 | 0.9910 | 0.9912 | 0.9914 | 0.9916 | 0.9917 | 0.9918 | 0.9919 | 0.9920 | 0.9921 | 0.9922 |
| 2.5 | 0.9923 | 0.9924 | 0.9925 | 0.9926 | 0.9927 | 0.9928 | 0.9929 | 0.9930 | 0.9931 | 0.9932 |
| 2.6 | 0.9933 | 0.9934 | 0.9935 | 0.9936 | 0.9937 | 0.9938 | 0.9939 | 0.9940 | 0.9941 | 0.9942 |
| 2.7 | 0.9943 | 0.9944 | 0.9945 | 0.9946 | 0.9947 | 0.9948 | 0.9949 | 0.9950 | 0.9951 | 0.9952 |
| 2.8 | 0.9953 | 0.9954 | 0.9955 | 0.9956 | 0.9957 | 0.9958 | 0.9959 | 0.9960 | 0.9961 | 0.9962 |
| 2.9 | 0.9963 | 0.9964 | 0.9965 | 0.9966 | 0.9967 | 0.9968 | 0.9969 | 0.9970 | 0.9971 | 0.9972 |
| 3.0 | 0.9973 | 0.9974 | 0.9975 | 0.9976 | 0.9977 | 0.9978 | 0.9979 | 0.9980 | 0.9981 | 0.9982 |
| 3.1 | 0.9983 | 0.9984 | 0.9985 | 0.9986 | 0.9987 | 0.9988 | 0.9989 | 0.9990 | 0.9991 | 0.9992 |
| 3.2 | 0.9993 | 0.9994 | 0.9995 | 0.9996 | 0.9997 | 0.9998 | 0.9999 | 1.0000 | 1.0000 | 1.0000 |

续表

| | | | | | | | | | | |
|------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| 0.87 | 1.126 39 | 1.131 34 | 1.134 896 | 1.140 467 | 1.145 565 | 1.150 349 | 1.155 221 | 1.160 120 | 1.165 047 | 1.170 002 |
| 0.88 | 1.171 587 | 1.180 001 | 1.185 044 | 1.190 118 | 1.195 223 | 1.200 359 | 1.205 327 | 1.210 727 | 1.215 960 | 1.221 227 |
| 0.89 | 1.226 26 | 1.231 864 | 1.237 23 | 1.242 641 | 1.248 085 | 1.253 565 | 1.259 384 | 1.264 641 | 1.270 238 | 1.275 874 |
| | 1.281 772 | 1.287 37 | 1.295 037 | 1.298 837 | 1.304 | 1.310 779 | 1.316 519 | 1.322 505 | 1.328 735 | 1.334 622 |
| | 1.34 775 | 1.346 939 | 1.353 74 | 1.359 463 | 1.36 | 1.372 204 | 1.378 659 | 1.385 172 | 1.391 744 | 1.398 477 |
| | 1.41 72 | 1.411 83 | 1.418 674 | 1.427 511 | 1.432 | 1.439 531 | 1.446 632 | 1.453 806 | 1.461 056 | 1.468 381 |
| | 1.479 | 1.483 28 | 1.488 83 | 1.498 71 | 1.506 | 1.514 102 | 1.522 036 | 1.530 068 | 1.538 199 | 1.546 181 |
| | 1.574 771 | 1.583 234 | 1.591 787 | 1.598 467 | 1.605 | 1.608 195 | 1.607 248 | 1.616 436 | 1.625 763 | 1.635 234 |
| | 1.644 854 | 1.651 628 | 1.664 561 | 1.674 65 | 1.684 | 1.687 598 | 1.705 343 | 1.716 886 | 1.727 934 | 1.739 136 |
| | 1.771 686 | 1.782 11 | 1.771 382 | 1.786 61 | 1.799 | 1.811 911 | 1.827 097 | 1.838 121 | 1.852 186 | 1.866 195 |
| | 1.887 91 | 1.89 698 | 1.911 36 | 1.925 87 | 1.933 | 1.93 664 | 1.977 368 | 1.995 893 | 2. 11 091 | 2.033 2 |
| | 2.075 719 | 2.074 855 | 2.086 927 | 2.12 72 | 2.111 | 2.170 090 | 2.197 286 | 2.226 212 | 2.257 129 | 2.290 368 |
| | 2.326 335 | 2.355 618 | 2.388 916 | 2.437 | 2.432 | 2.47 829 | 2.62 070 | 2.747 781 | 2.878 162 | 3.009 23 |

注:本表对正态概率给出正态分布的分位数 Z_p 。

例:对 $p = 0.9$, $Z_p = 1.644 871$ 。

附表 2 相关系数检验表

| 自由度 | 显著性水平 α | | | | | | | | | |
|-----|----------------|--------|--------|--------|--------|--------|--------|--------|---------|---------|
| | 0.10 | 0.05 | 0.025 | 0.01 | 0.005 | 0.001 | 0.0005 | 0.0001 | 0.00005 | 0.00001 |
| 1 | 0.6628 | 0.7081 | 0.7501 | 0.8783 | 0.9000 | 0.9500 | 0.9600 | 0.9750 | 0.9850 | 0.9900 |
| 2 | 0.6314 | 0.6757 | 0.7173 | 0.8538 | 0.8745 | 0.9250 | 0.9370 | 0.9525 | 0.9630 | 0.9670 |
| 3 | 0.6103 | 0.6538 | 0.6951 | 0.8334 | 0.8538 | 0.9050 | 0.9170 | 0.9325 | 0.9430 | 0.9470 |
| 4 | 0.5965 | 0.6391 | 0.6799 | 0.8191 | 0.8391 | 0.8900 | 0.9020 | 0.9175 | 0.9280 | 0.9320 |
| 5 | 0.5876 | 0.6293 | 0.6697 | 0.8114 | 0.8314 | 0.8825 | 0.8945 | 0.9100 | 0.9205 | 0.9245 |
| 6 | 0.5817 | 0.6225 | 0.6627 | 0.8061 | 0.8261 | 0.8775 | 0.8895 | 0.9050 | 0.9155 | 0.9195 |
| 7 | 0.5777 | 0.6176 | 0.6577 | 0.8021 | 0.8221 | 0.8735 | 0.8855 | 0.9010 | 0.9115 | 0.9155 |
| 8 | 0.5747 | 0.6146 | 0.6547 | 0.8000 | 0.8200 | 0.8715 | 0.8835 | 0.8990 | 0.9095 | 0.9135 |
| 9 | 0.5725 | 0.6124 | 0.6525 | 0.7985 | 0.8185 | 0.8700 | 0.8820 | 0.8975 | 0.9080 | 0.9120 |
| 10 | 0.5708 | 0.6107 | 0.6508 | 0.7973 | 0.8173 | 0.8688 | 0.8808 | 0.8963 | 0.9068 | 0.9108 |
| 11 | 0.5695 | 0.6093 | 0.6495 | 0.7963 | 0.8163 | 0.8678 | 0.8798 | 0.8953 | 0.9058 | 0.9098 |
| 12 | 0.5684 | 0.6082 | 0.6484 | 0.7955 | 0.8155 | 0.8670 | 0.8790 | 0.8945 | 0.9050 | 0.9090 |
| 13 | 0.5675 | 0.6072 | 0.6475 | 0.7948 | 0.8148 | 0.8663 | 0.8783 | 0.8938 | 0.9043 | 0.9083 |
| 14 | 0.5668 | 0.6063 | 0.6468 | 0.7942 | 0.8142 | 0.8657 | 0.8777 | 0.8932 | 0.9037 | 0.9077 |
| 15 | 0.5662 | 0.6055 | 0.6462 | 0.7937 | 0.8137 | 0.8652 | 0.8772 | 0.8927 | 0.9032 | 0.9072 |
| 16 | 0.5657 | 0.6048 | 0.6457 | 0.7933 | 0.8133 | 0.8647 | 0.8767 | 0.8922 | 0.9027 | 0.9067 |
| 17 | 0.5653 | 0.6042 | 0.6453 | 0.7930 | 0.8130 | 0.8643 | 0.8763 | 0.8918 | 0.9023 | 0.9063 |
| 18 | 0.5649 | 0.6037 | 0.6449 | 0.7927 | 0.8127 | 0.8640 | 0.8760 | 0.8915 | 0.9020 | 0.9060 |
| 19 | 0.5646 | 0.6033 | 0.6446 | 0.7925 | 0.8125 | 0.8637 | 0.8757 | 0.8912 | 0.9017 | 0.9057 |
| 20 | 0.5643 | 0.6030 | 0.6443 | 0.7923 | 0.8123 | 0.8635 | 0.8755 | 0.8910 | 0.9015 | 0.9055 |
| 21 | 0.5641 | 0.6027 | 0.6441 | 0.7921 | 0.8121 | 0.8633 | 0.8753 | 0.8908 | 0.9013 | 0.9053 |
| 22 | 0.5639 | 0.6025 | 0.6439 | 0.7920 | 0.8120 | 0.8632 | 0.8752 | 0.8907 | 0.9012 | 0.9052 |
| 23 | 0.5637 | 0.6023 | 0.6437 | 0.7919 | 0.8119 | 0.8631 | 0.8751 | 0.8906 | 0.9011 | 0.9051 |
| 24 | 0.5636 | 0.6022 | 0.6436 | 0.7918 | 0.8118 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 25 | 0.5635 | 0.6021 | 0.6435 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 26 | 0.5634 | 0.6020 | 0.6434 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 27 | 0.5634 | 0.6020 | 0.6434 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 28 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 29 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 30 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 31 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 32 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 33 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 34 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 35 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 36 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 37 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 38 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 39 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 40 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 41 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 42 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 43 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 44 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 45 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 46 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 47 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 48 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 49 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 50 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 51 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 52 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 53 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 54 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 55 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 56 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 57 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 58 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 59 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 60 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 61 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 62 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 63 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 64 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 65 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 66 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 67 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 68 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 69 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 70 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 71 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 72 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 73 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 74 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 75 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 76 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 77 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 78 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 79 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 80 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |
| 81 | 0.5633 | 0.6019 | 0.6433 | 0.7917 | 0.8117 | 0.8630 | 0.8750 | 0.8905 | 0.9010 | 0.9050 |

附表 3 χ^2 分布临界值表例如:自由度 $n=20$, $P(\chi^2 > 34.17) = 0.025$.

| n | $\alpha=0.995$ | $\alpha=0.990$ | $\alpha=0.975$ | $\alpha=0.950$ | $\alpha=0.900$ | $\alpha=0.025$ | $\alpha=0.010$ | $\alpha=0.005$ |
|-----|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| 1 | 0.000 939 3 | 0.000 157 | 0.000 982 | 0.003 93 | 3.841 | 5.024 | 6.635 | 7.879 |
| 2 | 0.010 | 0.020 1 | 0.050 6 | 0.103 | 5.991 | 7.378 | 9.210 | 10.579 |
| 3 | 0.072 | 0.115 | 0.216 | 0.352 | 7.815 | 9.348 | 11.345 | 12.838 |
| 4 | 0.207 | 0.297 | 0.484 | 0.711 | 9.488 | 11.143 | 13.277 | 14.860 |
| 5 | 0.412 | 0.554 | 0.831 | 1.145 | 11.070 | 12.832 | 15.086 | 16.750 |
| 6 | 0.676 | 0.872 | 1.237 | 1.635 | 12.592 | 14.449 | 16.812 | 18.548 |
| 7 | 0.989 | 1.239 | 1.690 | 2.167 | 14.067 | 16.013 | 18.475 | 20.278 |
| 8 | 1.344 | 1.646 | 2.180 | 2.733 | 15.507 | 17.535 | 20.090 | 21.955 |
| 9 | 1.735 | 2.088 | 2.700 | 3.325 | 16.919 | 19.023 | 21.666 | 23.589 |
| 10 | 2.156 | 2.558 | 3.247 | 3.940 | 18.307 | 20.483 | 23.209 | 25.188 |
| 11 | 2.603 | 3.053 | 3.816 | 4.575 | 19.675 | 21.920 | 24.725 | 26.757 |
| 12 | 3.074 | 3.571 | 4.404 | 5.226 | 21.026 | 23.337 | 26.217 | 28.306 |
| 13 | 3.565 | 4.107 | 5.009 | 5.892 | 22.362 | 24.736 | 27.688 | 29.819 |
| 14 | 4.075 | 4.660 | 5.629 | 6.571 | 23.685 | 26.119 | 29.141 | 31.319 |
| 15 | 4.601 | 5.229 | 6.262 | 7.261 | 24.996 | 27.488 | 30.578 | 32.801 |
| 16 | 5.142 | 5.812 | 6.908 | 7.962 | 26.296 | 28.845 | 32.000 | 34.267 |
| 17 | 5.697 | 6.408 | 7.564 | 8.672 | 27.587 | 30.191 | 33.409 | 35.718 |
| 18 | 6.265 | 7.015 | 8.231 | 9.390 | 28.869 | 31.526 | 34.805 | 37.156 |
| 19 | 6.844 | 7.633 | 8.907 | 10.117 | 30.144 | 32.852 | 36.191 | 38.582 |
| 20 | 7.434 | 8.260 | 9.591 | 10.851 | 31.410 | 34.170 | 37.566 | 39.997 |
| 21 | 8.034 | 8.897 | 10.283 | 11.591 | 32.671 | 35.479 | 38.932 | 41.401 |
| 22 | 8.643 | 9.542 | 10.982 | 12.338 | 33.924 | 36.781 | 40.289 | 42.796 |
| 23 | 9.260 | 10.196 | 11.689 | 13.091 | 35.172 | 38.076 | 41.638 | 44.181 |
| 24 | 9.886 | 10.856 | 12.401 | 13.848 | 36.415 | 39.364 | 42.980 | 45.558 |
| 25 | 10.520 | 11.524 | 13.120 | 14.611 | 37.652 | 40.646 | 44.314 | 46.928 |
| 26 | 11.160 | 12.198 | 13.844 | 15.379 | 38.885 | 41.923 | 45.642 | 48.290 |
| 27 | 11.808 | 12.879 | 14.573 | 16.151 | 40.113 | 43.194 | 46.963 | 49.645 |
| 28 | 12.461 | 13.565 | 15.308 | 16.928 | 41.337 | 44.461 | 48.278 | 50.993 |
| 29 | 13.121 | 14.256 | 16.047 | 17.708 | 42.557 | 45.722 | 49.588 | 52.336 |
| 30 | 13.787 | 14.953 | 16.791 | 18.493 | 43.773 | 46.979 | 50.892 | 53.672 |

附表 4 t 分布临界值表例如: 自由度 $n=20$, $P(t>1.725)=0.05$ 。

| | 25 | 0.20 | 0.15 | 0.10 | 0.05 | 0.025 | 0.01 | 0.005 | 0.000 5 |
|----------|-------|-------|-------|-------|-------|--------|--------|--------|---------|
| 1 | 0.100 | 1.376 | 1.963 | 3.076 | 6.314 | 12.706 | 31.821 | 63.657 | 636.619 |
| 2 | 0.816 | 1.061 | 1.386 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 | 31.598 |
| 3 | 0.765 | 0.978 | 1.250 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 12.941 |
| 4 | 0.741 | 0.941 | 1.190 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 8.610 |
| 5 | 0.727 | 0.920 | 1.156 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 6.859 |
| 6 | 0.718 | 0.906 | 1.134 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 | 5.959 |
| 7 | 0.711 | 0.896 | 1.119 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 | 5.405 |
| 8 | 0.706 | 0.889 | 1.108 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 | 5.041 |
| 9 | 0.703 | 0.883 | 1.100 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 | 4.781 |
| 10 | 0.700 | 0.879 | 1.093 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 4.587 |
| 11 | 0.697 | 0.876 | 1.088 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 | 4.437 |
| 12 | 0.695 | 0.873 | 1.083 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 | 4.318 |
| 13 | 0.694 | 0.870 | 1.079 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 | 4.221 |
| 14 | 0.692 | 0.868 | 1.076 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 | 4.140 |
| 15 | 0.691 | 0.866 | 1.074 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 | 4.073 |
| 16 | 0.690 | 0.865 | 1.071 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 | 4.015 |
| 17 | 0.689 | 0.863 | 1.069 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 | 3.965 |
| 18 | 0.688 | 0.862 | 1.067 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 | 3.922 |
| 19 | 0.688 | 0.861 | 1.066 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 | 3.883 |
| 20 | 0.687 | 0.860 | 1.064 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 | 3.850 |
| 21 | 0.686 | 0.859 | 1.063 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 | 3.819 |
| 22 | 0.686 | 0.858 | 1.061 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 | 3.792 |
| 23 | 0.685 | 0.858 | 1.060 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 | 3.767 |
| 24 | 0.685 | 0.857 | 1.059 | 1.318 | 1.711 | 2.064 | 2.492 | 2.397 | 3.745 |
| 25 | 0.684 | 0.856 | 1.058 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 | 3.725 |
| 26 | 0.684 | 0.856 | 1.058 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 | 3.707 |
| 27 | 0.684 | 0.855 | 1.057 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 | 3.690 |
| 28 | 0.683 | 0.855 | 1.056 | 1.313 | 1.701 | 2.048 | 2.467 | 2.733 | 3.674 |
| 29 | 0.683 | 0.854 | 1.055 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 | 3.659 |
| 30 | 0.683 | 0.854 | 1.055 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 | 3.646 |
| 40 | 0.681 | 0.851 | 1.050 | 1.303 | 1.684 | 2.021 | 2.423 | 2.704 | 3.551 |
| 60 | 0.679 | 0.848 | 1.046 | 1.296 | 1.671 | 2.000 | 2.390 | 2.660 | 3.460 |
| 120 | 0.677 | 0.845 | 1.041 | 1.289 | 1.658 | 1.980 | 2.358 | 2.617 | 3.373 |
| ∞ | 0.674 | 0.842 | 1.036 | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 | 3.291 |

续表

| m_1 | m_2 | 分子的自由度 | | | | | | | | | | | |
|----------------------------|-------|--------|------|------|------|------|------|------|------|------|------|------|------|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 分
母
的
自
由
度 | 23 | 4.28 | 3.42 | 3.03 | 2.80 | 2.64 | 2.53 | 2.45 | 2.38 | 2.32 | 2.28 | 2.24 | 2.20 |
| | | 7.88 | 5.66 | 4.76 | 4.26 | 3.94 | 3.71 | 3.54 | 3.41 | 3.30 | 3.21 | 3.14 | 3.07 |
| | 24 | 4.26 | 3.40 | 3.01 | 2.78 | 2.62 | 2.51 | 2.43 | 2.36 | 2.30 | 2.26 | 2.22 | 2.18 |
| | | 7.82 | 5.61 | 4.72 | 4.22 | 3.90 | 3.67 | 3.50 | 3.36 | 3.25 | 3.17 | 3.09 | 3.03 |
| | 25 | 4.24 | 3.38 | 2.99 | 2.76 | 2.60 | 2.49 | 2.41 | 2.32 | 2.28 | 2.24 | 2.20 | 2.16 |
| | | 7.77 | 5.57 | 4.68 | 4.18 | 3.86 | 3.63 | 3.46 | 3.34 | 3.21 | 3.13 | 3.05 | 2.99 |
| | 26 | 4.22 | 3.37 | 2.98 | 2.74 | 2.59 | 2.47 | 2.39 | 2.32 | 2.27 | 2.22 | 2.18 | 2.15 |
| | | 7.72 | 5.53 | 4.64 | 4.14 | 3.82 | 3.59 | 3.42 | 3.29 | 3.17 | 3.09 | 3.02 | 2.96 |
| | 27 | 4.21 | 3.35 | 2.96 | 2.73 | 2.57 | 2.46 | 2.37 | 2.30 | 2.25 | 2.20 | 2.16 | 2.13 |
| | | 7.68 | 5.49 | 4.60 | 4.11 | 3.79 | 3.56 | 3.39 | 3.26 | 3.14 | 3.06 | 2.98 | 2.93 |
| | 28 | 4.20 | 3.34 | 2.95 | 2.71 | 2.56 | 2.44 | 2.36 | 2.29 | 2.24 | 2.19 | 2.15 | 2.12 |
| | | 7.64 | 5.45 | 4.57 | 4.07 | 3.76 | 3.53 | 3.36 | 3.23 | 3.11 | 3.03 | 2.95 | 2.90 |
| | 29 | 4.18 | 3.33 | 2.93 | 2.70 | 2.54 | 2.43 | 2.35 | 2.28 | 2.22 | 2.18 | 2.14 | 2.10 |
| | | 7.60 | 5.42 | 4.54 | 4.04 | 3.73 | 3.50 | 3.33 | 3.20 | 3.08 | 3.00 | 2.92 | 2.87 |
| | 30 | 4.17 | 3.32 | 2.92 | 2.69 | 2.53 | 2.42 | 2.34 | 2.27 | 2.21 | 2.16 | 2.12 | 2.09 |
| | | 7.56 | 5.39 | 4.51 | 4.02 | 3.70 | 3.47 | 3.30 | 3.17 | 3.06 | 2.98 | 2.90 | 2.84 |
| | 32 | 4.15 | 3.30 | 2.90 | 2.67 | 2.51 | 2.40 | 2.32 | 2.25 | 2.19 | 2.14 | 2.10 | 2.07 |
| | | 7.50 | 5.34 | 4.46 | 3.97 | 3.66 | 3.42 | 3.25 | 3.12 | 3.01 | 2.94 | 2.86 | 2.80 |
| | 34 | 4.13 | 3.28 | 2.88 | 2.65 | 2.49 | 2.38 | 2.30 | 2.23 | 2.17 | 2.12 | 2.08 | 2.50 |
| | | 7.44 | 5.29 | 4.42 | 3.93 | 3.61 | 3.38 | 3.21 | 3.08 | 2.97 | 2.89 | 2.82 | 2.76 |
| | 36 | 4.11 | 3.26 | 2.86 | 2.63 | 2.48 | 2.36 | 2.28 | 2.21 | 2.15 | 2.10 | 2.06 | 2.03 |
| | | 7.39 | 5.25 | 4.38 | 3.80 | 3.58 | 3.35 | 3.18 | 3.04 | 2.94 | 2.86 | 2.78 | 2.72 |
| | 38 | 4.10 | 3.25 | 2.85 | 2.62 | 2.46 | 2.35 | 2.26 | 2.19 | 2.14 | 2.09 | 2.05 | 2.02 |
| | | 7.35 | 5.21 | 4.34 | 3.86 | 3.54 | 3.32 | 3.15 | 3.02 | 2.91 | 2.82 | 2.75 | 2.69 |
| | 40 | 4.08 | 3.23 | 2.84 | 2.61 | 2.45 | 2.34 | 2.25 | 2.18 | 2.12 | 2.07 | 2.04 | 2.00 |
| | | 7.31 | 5.18 | 4.31 | 3.83 | 3.51 | 3.29 | 3.12 | 2.99 | 2.88 | 2.80 | 2.73 | 2.66 |
| | 42 | 4.07 | 3.22 | 2.83 | 2.59 | 2.44 | 2.32 | 2.24 | 2.17 | 2.11 | 2.06 | 2.02 | 1.99 |
| | | 7.27 | 5.15 | 4.29 | 3.80 | 3.49 | 3.26 | 3.10 | 2.96 | 2.86 | 2.77 | 2.70 | 2.64 |
| | 44 | 4.06 | 3.21 | 2.82 | 2.58 | 2.43 | 2.34 | 2.23 | 2.16 | 2.10 | 2.05 | 2.01 | 1.98 |
| | | 7.24 | 5.12 | 4.26 | 3.78 | 3.46 | 3.24 | 3.07 | 2.94 | 2.84 | 2.75 | 2.68 | 2.62 |
| | 46 | 4.05 | 3.20 | 2.81 | 2.57 | 2.42 | 2.30 | 2.22 | 2.14 | 2.09 | 2.04 | 2.00 | 1.97 |
| | | 7.21 | 5.10 | 4.24 | 3.76 | 3.44 | 3.22 | 3.05 | 2.92 | 2.82 | 2.73 | 2.66 | 2.60 |
| | 48 | 4.04 | 3.19 | 2.80 | 2.56 | 2.41 | 2.30 | 2.21 | 2.14 | 2.08 | 2.03 | 1.99 | 1.96 |
| | | 7.19 | 5.08 | 4.22 | 3.74 | 3.42 | 3.20 | 3.04 | 2.90 | 2.80 | 2.71 | 2.64 | 2.58 |
| | 50 | 4.03 | 3.18 | 2.79 | 2.56 | 2.40 | 2.29 | 2.20 | 2.13 | 2.07 | 2.02 | 1.98 | 1.95 |
| | | 7.17 | 5.06 | 4.20 | 3.72 | 3.41 | 3.18 | 3.02 | 2.88 | 2.78 | 2.70 | 2.62 | 2.56 |
| | 55 | 4.02 | 3.17 | 2.78 | 2.54 | 2.38 | 2.27 | 2.18 | 2.11 | 2.05 | 2.00 | 1.97 | 1.93 |
| | | 7.12 | 5.01 | 4.16 | 3.68 | 3.37 | 3.15 | 2.98 | 2.85 | 2.75 | 2.66 | 2.59 | 2.53 |
| | 60 | 4.00 | 3.15 | 2.76 | 2.52 | 2.37 | 2.25 | 2.17 | 2.10 | 2.04 | 1.99 | 1.95 | 1.92 |
| | | 7.08 | 4.98 | 4.13 | 3.65 | 3.34 | 3.12 | 2.95 | 2.82 | 2.72 | 2.63 | 2.56 | 2.50 |
| | 65 | 3.99 | 3.14 | 2.75 | 2.51 | 2.36 | 2.24 | 2.15 | 2.08 | 2.02 | 1.98 | 1.94 | 1.90 |
| | | 7.04 | 4.95 | 4.10 | 3.62 | 3.31 | 3.09 | 2.93 | 2.79 | 2.70 | 2.61 | 2.54 | 2.47 |
| | 70 | 3.98 | 3.13 | 2.74 | 2.50 | 2.35 | 2.23 | 2.14 | 2.07 | 2.01 | 1.97 | 1.93 | 1.89 |
| | | 7.01 | 4.92 | 4.08 | 3.60 | 3.29 | 3.07 | 2.91 | 2.77 | 2.67 | 2.59 | 2.51 | 2.45 |
| | 80 | 3.96 | 3.11 | 2.72 | 2.48 | 2.33 | 2.21 | 2.12 | 2.05 | 1.99 | 1.95 | 1.91 | 1.88 |
| | | 6.96 | 4.88 | 4.04 | 3.56 | 3.25 | 3.04 | 2.87 | 2.74 | 2.64 | 2.55 | 2.48 | 2.41 |
| | 100 | 3.94 | 3.09 | 2.70 | 2.46 | 2.30 | 2.19 | 2.10 | 2.03 | 1.97 | 1.92 | 1.88 | 1.85 |
| | | 6.90 | 4.82 | 3.98 | 3.51 | 3.20 | 2.99 | 2.82 | 2.69 | 2.59 | 2.51 | 2.43 | 2.36 |
| | 125 | 3.92 | 3.07 | 2.68 | 2.44 | 2.29 | 2.17 | 2.08 | 2.01 | 1.95 | 1.90 | 1.86 | 1.83 |
| | | 6.84 | 4.78 | 3.94 | 3.47 | 3.17 | 2.95 | 2.79 | 2.65 | 2.56 | 2.47 | 2.40 | 2.33 |
| | 150 | 3.91 | 3.06 | 2.67 | 2.43 | 2.27 | 2.16 | 2.07 | 2.00 | 1.94 | 1.89 | 1.85 | 1.82 |
| | | 6.81 | 4.75 | 3.91 | 3.44 | 3.14 | 2.92 | 2.76 | 2.62 | 2.53 | 2.44 | 2.37 | 2.30 |

续表

| n_1 | | 分子的自由度 | | | | | | | | | | | |
|--------|------|--------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| n_2 | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 分母的自由度 | 200 | 3.89 | 3.04 | 2.65 | 2.41 | 2.26 | 2.14 | 2.05 | 1.98 | 1.92 | 1.87 | 1.83 | 1.80 |
| | 400 | 6.76 | 4.71 | 3.88 | 3.41 | 3.11 | 2.90 | 2.73 | 2.60 | 2.50 | 2.41 | 2.34 | 2.28 |
| | 600 | 3.86 | 3.02 | 2.62 | 2.39 | 2.23 | 2.12 | 2.03 | 1.96 | 1.90 | 1.85 | 1.81 | 1.78 |
| | 800 | 5.70 | 4.66 | 3.83 | 3.36 | 3.06 | 2.85 | 2.69 | 2.55 | 2.46 | 2.37 | 2.29 | 2.23 |
| | 1000 | 3.85 | 3.00 | 2.61 | 2.38 | 2.22 | 2.10 | 2.02 | 1.95 | 1.89 | 1.84 | 1.80 | 1.76 |
| | ∞ | 6.66 | 4.62 | 3.80 | 3.34 | 3.04 | 2.82 | 2.66 | 2.53 | 2.43 | 2.34 | 2.26 | 2.20 |
| | | 3.84 | 2.99 | 2.60 | 2.37 | 2.21 | 2.09 | 2.01 | 1.94 | 1.88 | 1.83 | 1.79 | 1.75 |
| | | 6.64 | 4.60 | 3.78 | 3.32 | 3.02 | 2.80 | 2.64 | 2.51 | 2.41 | 2.32 | 2.24 | 2.18 |
| | | | | | | | | | | | | | |
| | | | | | | | | | | | | | |
| n_1 | | 分子的自由度 | | | | | | | | | | | |
| n_2 | | 14 | 16 | 20 | 24 | 30 | 40 | 50 | 75 | 100 | 200 | 500 | ∞ |
| 分母的自由度 | 1 | 245 | 246 | 248 | 249 | 250 | 251 | 252 | 253 | 253 | 254 | 254 | 254 |
| | 2 | 6 142 | 6 169 | 6 208 | 6 234 | 6 258 | 6 286 | 6 302 | 6 323 | 6 334 | 6 352 | 6 361 | 6 366 |
| | 3 | 19.42 | 19.43 | 19.44 | 19.45 | 19.46 | 19.47 | 19.47 | 19.48 | 19.49 | 19.49 | 19.50 | 19.50 |
| | 4 | 99.43 | 99.44 | 99.45 | 99.46 | 99.47 | 99.48 | 99.48 | 99.49 | 99.49 | 99.49 | 99.50 | 99.50 |
| | 5 | 8.71 | 8.69 | 8.66 | 8.64 | 8.62 | 8.60 | 8.58 | 8.57 | 8.56 | 8.54 | 8.54 | 8.53 |
| | 6 | 26.92 | 26.83 | 26.69 | 26.60 | 26.50 | 26.41 | 26.35 | 26.27 | 26.23 | 26.18 | 26.14 | 26.12 |
| | 7 | 5.87 | 5.84 | 5.80 | 5.77 | 5.74 | 5.71 | 5.70 | 5.68 | 5.66 | 5.65 | 5.64 | 5.63 |
| | 8 | 14.24 | 14.15 | 14.02 | 13.93 | 13.83 | 13.74 | 13.69 | 13.61 | 13.57 | 13.52 | 13.48 | 13.46 |
| | 9 | 4.64 | 4.60 | 4.56 | 4.53 | 4.50 | 4.46 | 4.44 | 4.42 | 4.40 | 4.38 | 4.37 | 4.36 |
| | 10 | 9.77 | 9.68 | 9.55 | 9.47 | 9.38 | 9.29 | 9.24 | 9.17 | 9.13 | 9.07 | 9.04 | 9.02 |
| | 11 | 3.96 | 3.92 | 3.87 | 3.84 | 3.81 | 3.77 | 3.75 | 3.72 | 3.71 | 3.69 | 3.68 | 3.67 |
| | 12 | 7.60 | 7.52 | 7.39 | 7.31 | 7.23 | 7.14 | 7.09 | 7.02 | 6.99 | 6.94 | 6.90 | 6.88 |
| | 13 | 3.52 | 3.49 | 3.44 | 3.41 | 3.38 | 3.34 | 3.32 | 3.29 | 3.28 | 3.25 | 3.24 | 3.23 |
| | 14 | 6.35 | 6.27 | 6.15 | 6.07 | 5.98 | 5.90 | 5.85 | 5.78 | 5.75 | 5.70 | 5.67 | 5.65 |
| | 15 | 3.23 | 3.20 | 3.15 | 3.12 | 3.08 | 3.05 | 3.03 | 3.00 | 2.98 | 2.96 | 2.94 | 2.93 |
| | 16 | 5.56 | 5.48 | 5.36 | 5.28 | 5.20 | 5.11 | 5.06 | 5.00 | 4.96 | 4.91 | 4.88 | 4.86 |
| | 17 | 3.02 | 2.98 | 2.93 | 2.90 | 2.86 | 2.82 | 2.80 | 2.77 | 2.76 | 2.73 | 2.72 | 2.71 |
| | 18 | 5.00 | 4.92 | 4.80 | 4.73 | 4.64 | 4.56 | 4.51 | 4.45 | 4.41 | 4.36 | 4.33 | 4.31 |
| | 19 | 2.86 | 2.82 | 2.77 | 2.74 | 2.70 | 2.67 | 2.64 | 2.61 | 2.59 | 2.56 | 2.55 | 2.54 |
| | 20 | 4.60 | 4.52 | 4.41 | 4.33 | 4.25 | 4.17 | 4.12 | 4.05 | 4.01 | 3.96 | 3.93 | 3.94 |
| | 21 | 2.74 | 2.70 | 2.65 | 2.61 | 2.57 | 2.53 | 2.50 | 2.47 | 2.45 | 2.42 | 2.41 | 2.40 |
| | 22 | 4.29 | 4.21 | 4.10 | 4.02 | 3.94 | 3.86 | 3.80 | 3.74 | 3.70 | 3.66 | 3.62 | 3.60 |
| | 23 | 2.64 | 2.60 | 2.54 | 2.50 | 2.46 | 2.42 | 2.40 | 2.36 | 2.35 | 2.32 | 2.31 | 2.30 |
| | 24 | 4.05 | 3.98 | 3.86 | 3.78 | 3.70 | 3.61 | 3.56 | 3.49 | 3.46 | 3.41 | 3.38 | 3.36 |
| | 25 | 2.55 | 2.51 | 2.46 | 2.42 | 2.38 | 2.34 | 2.32 | 2.28 | 2.26 | 2.24 | 2.22 | 2.21 |
| | 26 | 3.85 | 3.78 | 3.67 | 3.59 | 3.51 | 3.42 | 3.37 | 3.30 | 3.27 | 3.21 | 3.18 | 3.16 |
| | 27 | 2.48 | 2.44 | 2.39 | 2.35 | 2.31 | 2.27 | 2.24 | 2.21 | 2.19 | 2.16 | 2.14 | 2.13 |
| | 28 | 3.70 | 3.62 | 3.51 | 3.43 | 3.34 | 3.26 | 3.21 | 3.14 | 3.11 | 3.06 | 3.02 | 3.00 |
| | 29 | 2.48 | 2.39 | 2.33 | 2.29 | 2.25 | 2.21 | 2.18 | 2.15 | 2.12 | 2.10 | 2.08 | 2.07 |
| | 30 | 3.56 | 3.48 | 3.36 | 3.29 | 3.20 | 3.12 | 3.07 | 3.00 | 2.97 | 2.92 | 2.89 | 2.87 |
| | 31 | 2.37 | 2.33 | 2.28 | 2.24 | 2.20 | 2.16 | 2.13 | 2.09 | 2.07 | 2.04 | 2.02 | 2.01 |
| | 32 | 3.45 | 3.37 | 3.25 | 3.18 | 3.10 | 3.01 | 2.96 | 2.89 | 2.86 | 2.80 | 2.77 | 2.75 |
| | 33 | 2.33 | 2.29 | 2.23 | 2.19 | 2.15 | 2.11 | 2.08 | 2.04 | 2.02 | 1.99 | 1.97 | 1.96 |
| | 34 | 3.35 | 3.27 | 3.16 | 3.08 | 3.00 | 2.92 | 2.86 | 2.79 | 2.76 | 2.70 | 2.67 | 2.65 |
| | 35 | 2.29 | 2.25 | 2.19 | 2.15 | 2.11 | 2.07 | 2.04 | 2.00 | 1.98 | 1.95 | 1.93 | 1.92 |
| | 36 | 3.27 | 3.19 | 3.07 | 3.00 | 2.91 | 2.83 | 2.78 | 2.71 | 2.68 | 2.62 | 2.59 | 2.57 |
| | 37 | 2.26 | 2.21 | 2.15 | 2.11 | 2.07 | 2.02 | 2.00 | 1.96 | 1.94 | 1.91 | 1.90 | 1.88 |
| | 38 | 3.19 | 3.12 | 3.00 | 2.92 | 2.84 | 2.76 | 2.70 | 2.63 | 2.60 | 2.54 | 2.51 | 2.49 |

续表

| n_1 | n_2 | 分子的自由度 | | | | | | | | | | | |
|----------------------------|-------|--------|------|------|------|------|------|------|------|------|------|------|----------|
| | | 14 | 16 | 20 | 24 | 30 | 40 | 50 | 75 | 100 | 200 | 500 | ∞ |
| 分
母
的
自
由
度 | 20 | 2.23 | 2.18 | 2.12 | 2.08 | 2.04 | 1.99 | 1.96 | 1.92 | 1.90 | 1.87 | 1.85 | 1.84 |
| | | 3.13 | 3.05 | 2.94 | 2.86 | 2.77 | 2.69 | 2.63 | 2.56 | 2.53 | 2.47 | 2.44 | 2.42 |
| | 21 | 2.20 | 2.15 | 2.09 | 2.05 | 2.00 | 1.96 | 1.93 | 1.89 | 1.87 | 1.84 | 1.82 | 1.81 |
| | | 3.07 | 2.99 | 2.88 | 2.80 | 2.72 | 2.63 | 2.58 | 2.51 | 2.47 | 2.42 | 2.38 | 2.36 |
| | 22 | 2.18 | 2.13 | 2.07 | 2.03 | 1.98 | 1.93 | 1.91 | 1.87 | 1.84 | 1.81 | 1.80 | 1.78 |
| | | 3.02 | 2.94 | 2.83 | 2.75 | 2.67 | 2.58 | 2.53 | 2.46 | 2.42 | 2.37 | 2.33 | 2.31 |
| | 23 | 2.14 | 2.10 | 2.04 | 2.00 | 1.96 | 1.91 | 1.88 | 1.84 | 1.82 | 1.79 | 1.77 | 1.76 |
| | | 2.97 | 2.89 | 2.78 | 2.79 | 2.62 | 2.53 | 2.48 | 2.41 | 2.37 | 2.32 | 2.28 | 2.26 |
| | 24 | 2.13 | 2.09 | 2.02 | 1.98 | 1.94 | 1.89 | 1.86 | 1.82 | 1.80 | 1.76 | 1.74 | 1.73 |
| | | 2.93 | 2.85 | 2.74 | 2.66 | 2.58 | 2.49 | 2.44 | 2.36 | 2.33 | 2.27 | 2.23 | 2.21 |
| | 25 | 2.11 | 2.06 | 2.00 | 1.96 | 1.92 | 1.87 | 1.84 | 1.80 | 1.77 | 1.74 | 1.72 | 1.71 |
| | | 2.89 | 2.81 | 2.70 | 2.62 | 2.54 | 2.45 | 2.40 | 2.32 | 2.29 | 2.23 | 2.19 | 2.17 |
| | 26 | 2.10 | 2.05 | 1.99 | 1.95 | 1.90 | 1.85 | 1.82 | 1.78 | 1.76 | 1.72 | 1.70 | 1.69 |
| | | 2.86 | 2.77 | 2.66 | 2.58 | 2.50 | 2.41 | 2.36 | 2.28 | 2.25 | 2.19 | 2.15 | 2.13 |
| | 27 | 2.08 | 2.03 | 1.97 | 1.93 | 1.88 | 1.84 | 1.80 | 1.76 | 1.74 | 1.71 | 1.68 | 1.67 |
| | | 2.83 | 2.74 | 2.63 | 2.55 | 2.47 | 2.38 | 2.33 | 2.25 | 2.21 | 2.16 | 2.12 | 2.10 |
| | 28 | 2.06 | 2.02 | 1.96 | 1.91 | 1.87 | 1.81 | 1.78 | 1.75 | 1.72 | 1.69 | 1.67 | 1.65 |
| | | 2.80 | 2.71 | 2.60 | 2.52 | 2.44 | 2.35 | 2.30 | 2.22 | 2.18 | 2.13 | 2.09 | 2.06 |
| | 29 | 2.05 | 2.00 | 1.94 | 1.90 | 1.85 | 1.80 | 1.77 | 1.73 | 1.71 | 1.68 | 1.65 | 1.64 |
| | | 2.77 | 2.68 | 2.57 | 2.49 | 2.41 | 2.32 | 2.27 | 2.19 | 2.15 | 2.10 | 2.06 | 2.03 |
| | 30 | 2.04 | 1.99 | 1.93 | 1.89 | 1.84 | 1.79 | 1.76 | 1.72 | 1.69 | 1.66 | 1.64 | 1.62 |
| | | 2.74 | 2.66 | 2.55 | 2.47 | 2.38 | 2.29 | 2.24 | 2.16 | 2.13 | 2.07 | 2.03 | 2.01 |
| | 32 | 2.02 | 1.97 | 1.91 | 1.86 | 1.82 | 1.76 | 1.74 | 1.69 | 1.67 | 1.64 | 1.61 | 1.59 |
| | | 2.70 | 2.62 | 2.51 | 2.42 | 2.34 | 2.25 | 2.20 | 2.12 | 2.08 | 2.02 | 1.98 | 1.96 |
| | 34 | 2.00 | 1.95 | 1.89 | 1.84 | 1.80 | 1.74 | 1.71 | 1.67 | 1.64 | 1.61 | 1.59 | 1.57 |
| | | 2.66 | 2.58 | 2.47 | 2.38 | 2.30 | 2.21 | 2.15 | 2.08 | 2.04 | 1.98 | 1.94 | 1.91 |
| | 36 | 1.98 | 1.93 | 1.87 | 1.82 | 1.78 | 1.72 | 1.69 | 1.65 | 1.62 | 1.59 | 1.56 | 1.55 |
| | | 2.62 | 2.54 | 2.43 | 2.35 | 2.26 | 2.17 | 2.12 | 2.04 | 2.00 | 1.94 | 1.90 | 1.87 |
| | 38 | 1.96 | 1.92 | 1.85 | 1.80 | 1.76 | 1.71 | 1.67 | 1.63 | 1.60 | 1.57 | 1.54 | 1.53 |
| | | 2.59 | 2.51 | 2.40 | 2.32 | 2.22 | 2.14 | 2.08 | 2.00 | 1.97 | 1.90 | 1.86 | 1.84 |
| | 40 | 1.95 | 1.90 | 1.84 | 1.79 | 1.74 | 1.69 | 1.66 | 1.61 | 1.59 | 1.55 | 1.53 | 1.51 |
| | | 2.56 | 2.49 | 2.37 | 2.29 | 2.20 | 2.11 | 2.05 | 1.97 | 1.94 | 1.88 | 1.84 | 1.81 |
| | 42 | 1.94 | 1.89 | 1.82 | 1.78 | 1.73 | 1.68 | 1.64 | 1.60 | 1.57 | 1.54 | 1.51 | 1.49 |
| | | 2.54 | 2.46 | 2.35 | 2.26 | 2.17 | 2.08 | 2.02 | 1.94 | 1.91 | 1.85 | 1.80 | 1.78 |
| | 44 | 1.92 | 1.88 | 1.81 | 1.76 | 1.72 | 1.66 | 1.63 | 1.58 | 1.56 | 1.52 | 1.50 | 1.48 |
| | | 2.52 | 2.44 | 2.32 | 2.24 | 2.15 | 2.06 | 2.00 | 1.92 | 1.88 | 1.82 | 1.78 | 1.75 |
| | 46 | 1.91 | 1.87 | 1.80 | 1.75 | 1.71 | 1.65 | 1.62 | 1.57 | 1.54 | 1.51 | 1.48 | 1.46 |
| | | 2.50 | 2.42 | 2.30 | 2.22 | 2.13 | 2.04 | 1.98 | 1.90 | 1.86 | 1.80 | 1.76 | 1.72 |
| | 48 | 1.90 | 1.86 | 1.79 | 1.74 | 1.70 | 1.64 | 1.61 | 1.56 | 1.53 | 1.50 | 1.47 | 1.45 |
| | | 2.48 | 2.40 | 2.28 | 2.20 | 2.11 | 2.02 | 1.96 | 1.88 | 1.84 | 1.78 | 1.73 | 1.70 |
| | 50 | 1.90 | 1.85 | 1.78 | 1.74 | 1.69 | 1.63 | 1.60 | 1.55 | 1.52 | 1.48 | 1.46 | 1.44 |
| | | 2.46 | 2.39 | 2.26 | 2.18 | 2.10 | 2.00 | 1.94 | 1.86 | 1.82 | 1.76 | 1.71 | 1.68 |
| | 55 | 1.88 | 1.83 | 1.76 | 1.72 | 1.67 | 1.61 | 1.58 | 1.52 | 1.50 | 1.46 | 1.43 | 1.41 |
| | | 2.43 | 2.35 | 2.23 | 2.15 | 2.06 | 1.96 | 1.90 | 1.82 | 1.78 | 1.71 | 1.66 | 1.64 |
| | 60 | 1.86 | 1.81 | 1.75 | 1.70 | 1.65 | 1.59 | 1.56 | 1.50 | 1.48 | 1.44 | 1.41 | 1.39 |
| | | 2.40 | 2.32 | 2.20 | 2.12 | 2.03 | 1.93 | 1.87 | 1.79 | 1.74 | 1.68 | 1.63 | 1.60 |
| | 65 | 1.85 | 1.80 | 1.73 | 1.68 | 1.63 | 1.57 | 1.54 | 1.49 | 1.46 | 1.42 | 1.39 | 1.37 |
| | | 2.37 | 2.30 | 2.18 | 2.09 | 2.00 | 1.90 | 1.84 | 1.76 | 1.71 | 1.64 | 1.60 | 1.56 |
| | 70 | 1.84 | 1.79 | 1.72 | 1.67 | 1.62 | 1.56 | 1.53 | 1.47 | 1.45 | 1.40 | 1.37 | 1.35 |
| | | 2.35 | 2.28 | 2.15 | 2.07 | 1.98 | 1.88 | 1.82 | 1.74 | 1.69 | 1.62 | 1.56 | 1.53 |
| | 80 | 1.82 | 1.77 | 1.70 | 1.65 | 1.60 | 1.54 | 1.51 | 1.45 | 1.42 | 1.38 | 1.35 | 1.32 |
| | | 2.32 | 2.24 | 2.11 | 2.03 | 1.94 | 1.84 | 1.78 | 1.70 | 1.65 | 1.57 | 1.52 | 1.49 |

续表

| $\begin{matrix} n_1 \\ n_2 \end{matrix}$ | | 分子的自由度 | | | | | | | | | | | |
|--|----------|--------|------|------|------|------|------|------|------|------|------|------|----------|
| | | 14 | 16 | 20 | 24 | 30 | 40 | 50 | 75 | 100 | 200 | 500 | ∞ |
| 分
母
的
自
由
度 | 100 | 1.79 | 1.75 | 1.68 | 1.63 | 1.57 | 1.51 | 1.48 | 1.42 | 1.39 | 1.34 | 1.30 | 1.28 |
| | | 2.26 | 2.19 | 2.06 | 1.98 | 1.89 | 1.79 | 1.73 | 1.64 | 1.59 | 1.51 | 1.46 | 1.43 |
| | 125 | 1.77 | 1.72 | 1.65 | 1.60 | 1.55 | 1.49 | 1.45 | 1.39 | 1.36 | 1.31 | 1.27 | 1.25 |
| | | 2.23 | 2.15 | 2.03 | 1.94 | 1.85 | 1.75 | 1.68 | 1.59 | 1.54 | 1.46 | 1.40 | 1.37 |
| | 150 | 1.76 | 1.71 | 1.64 | 1.59 | 1.54 | 1.47 | 1.44 | 1.37 | 1.34 | 1.29 | 1.25 | 1.22 |
| | | 2.20 | 2.12 | 2.00 | 1.91 | 1.83 | 1.72 | 1.66 | 1.56 | 1.51 | 1.43 | 1.37 | 1.33 |
| | 200 | 1.74 | 1.69 | 1.62 | 1.57 | 1.52 | 1.45 | 1.42 | 1.35 | 1.32 | 1.26 | 1.22 | 1.19 |
| | | 2.17 | 2.09 | 1.97 | 1.88 | 1.79 | 1.69 | 1.62 | 1.53 | 1.48 | 1.39 | 1.33 | 1.28 |
| | 400 | 1.72 | 1.67 | 1.60 | 1.54 | 1.49 | 1.42 | 1.38 | 1.32 | 1.28 | 1.22 | 1.16 | 1.13 |
| | | 2.12 | 2.04 | 1.92 | 1.84 | 1.74 | 1.64 | 1.57 | 1.47 | 1.42 | 1.32 | 1.24 | 1.19 |
| | 1 000 | 1.70 | 1.65 | 1.58 | 1.53 | 1.47 | 1.41 | 1.36 | 1.30 | 1.26 | 1.19 | 1.13 | 1.08 |
| | | 2.09 | 2.01 | 1.89 | 1.81 | 1.71 | 1.61 | 1.54 | 1.44 | 1.38 | 1.28 | 1.19 | 1.11 |
| | ∞ | 1.67 | 1.64 | 1.57 | 1.52 | 1.46 | 1.40 | 1.35 | 1.28 | 1.24 | 1.17 | 1.11 | 1.00 |
| | | 2.07 | 1.99 | 1.87 | 1.79 | 1.69 | 1.59 | 1.52 | 1.41 | 1.36 | 1.25 | 1.15 | 1.00 |